Correction TD1 – Question 1 à 5

Econométrie des variables qualitatives 1
Modèle logit binaire

Exercice : L'hypertension artérielle et ses facteurs explicatifs

Vous disposez pour 500 personnes de leur statut en termes d'hypertension (Press_arter) ainsi que des caractéristiques sur leur comportement. Les données sont disponibles dans la Base_pression_arterielle.xls sous Madoc.

- Press_arter : la pression systolique se répartit en 4 classes ordonnées

Classe	Modalités (millimètre de mercure, mmHg)
1 : Pression artérielle normale	< 140
2 : Hypertension artérielle de grade 1	[140-159[
3 : Hypertension artérielle de grade 2	[160-179[
4 : Hypertension artérielle de grade 3	≥ 179

Genre = 1 si la personne est un homme, 0 sinon
Fumer = 1 si la personne fume, 0 sinon
Sport = 1 si la personne pratique le sport de manière intensive, 0 sinon
Age : Age de la personne
Alcool =1 si la personne boit de l'alcool de manière excessive, 0 sinon
IMC : indice de masse corporelle
Stress = 1 si la personne est stressée, 0 sinon
Sel = 1 si l'alimentation de la personne est très salée, 0 sinon

Il existe également la variable Pression = 1 si la pression artérielle de la personne est supérieure à 140 (donc supérieure à la normale), 0 dans le cas contraire

Question 1 : Importer la base sous le logiciel R. Nommer la base : Pression

```
getwd()
setwd("C:/Users/travers-m/Desktop/Cours_2020_2021/Econometrie_variables_qualitatives_M1_EKAP_M2_CODEME/TD/Bases")
library(readxl)
Pression<-read_excel("Base_pression_arterielle.xls",sheet="Feuil1",col_names=TRUE)
```

<u>Question 2</u> : Vérifier la corrélation entre les différentes explicatives quantitatives. Qu'en concluez-vous ?

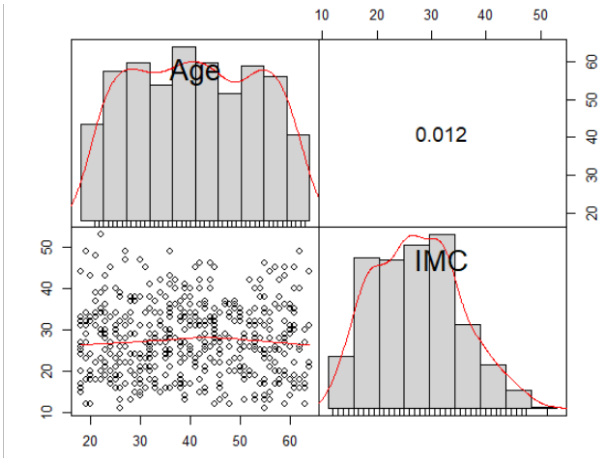
round(cor(Pression[,c("Age","IMC")],use="complete.obs",method = c("spearman")),3)

```
       Age   IMC
Age  1.000 0.012
IMC  0.012 1.000
```

➔ Il n'existe pas de corrélation significative importante entre les variables explicatives quantitatives. Attention si vous le faites avant le nettoyage de la base, il faudra refaire l'analyse des corrélations si la base est modifiée.

Ou :

library(PerformanceAnalytics)
mydata <- Pression[, c("Age","IMC")]
chart.Correlation(mydata, histogram=TRUE, pch=19,method = c("spearman"))



Pour les variables qualitatives, il faut tout d'abord les factoriser (cf. question 3)

<u>Question 3</u> : Transformer les variables de type qualitatif en facteur.

str(Pression)

'data.frame': 500 obs. of 9 variables:

$ Press_arter: num 1 1 2 1 1 1 1 1 1 1 ...
$ Genre : num 0 1 1 1 1 0 0 0 1 1 ...
$ Fumer : num 0 1 1 0 0 0 0 0 0 1 ...
$ Sport : num 1 0 0 0 0 1 0 1 0 1 ...
$ Age : num 60 55 18 19 58 55 22 52 46 38 ...
$ Alcool : num 0 0 0 1 1 0 1 0 1 0 ...
$ IMC : num 35 17 26 49 25 25 30 19 13 22 ...
$ Stress : num 0 0 1 1 0 0 1 0 0 0 ...
$ Sel : num 0 0 0 1 0 0 0 1 0 0 ...

Il faut transformer les variables Genre, Fumer, Sport, Alcool, Stress et Sel

Pression$Genre<-as.factor(Pression$Genre)
Pression$Fumer<-as.factor(Pression$Fumer)
Pression$Sport<-as.factor(Pression$Sport)
Pression$Alcool<-as.factor(Pression$Alcool)
Pression$Stress<-as.factor(Pression$Stress)
Pression$Sel<-as.factor(Pression$Sel)

<u>Deuxième méthode possible pour factoriser les variables :</u>

#Si vous prenez la deuxième base sous Madoc avec la colonne Obs
#Indiquer dans la ligne de commande ci-dessous le numéro des colonnes qui sont à mettre en facteur
Pression[,c(3,4,5,6,8,10,11)]=lapply(Pression[,c(3,4,5,6,8,10,11)],as.factor)
#Indiquer dans la ligne de commande ci-dessous le numéro des colonnes qui sont à mettre en numérique (y compris Press_arter) qui sera ensuite mise en variable ordonnée
Pression[,c(1,2,7,9)]=lapply(Pression[,c(1,2,7,9)],as.numeric)

Il faut également transformer la variable Press_arter en variable ordonnée :

Pression$Press_arter<-ordered(Pression$Press_arter)

str(Pression)

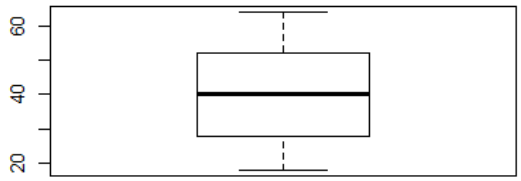
'data.frame': 500 obs. of 9 variables:
$ Press_arter: Ord.factor w/ 4 levels "1"<"2"<"3"<"4": 1 1 2 1 1 1 1 1 1 1 ...
$ Genre : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 1 1 2 2 ...
$ Fumer : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 2 ...
$ Sport : Factor w/ 2 levels "0","1": 2 1 1 1 1 2 1 2 1 2 ...
$ Age : num 60 55 18 19 58 55 22 52 46 38 ...
$ Alcool : Factor w/ 2 levels "0","1": 1 1 1 2 2 1 2 1 2 1 ...
$ IMC : num 35 17 26 49 25 25 30 19 13 22 ...
$ Stress : Factor w/ 2 levels "0","1": 1 1 2 2 1 1 2 1 1 1 ...
$ Sel : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 1 ...

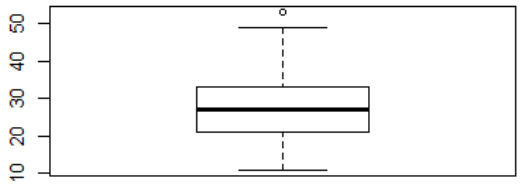
<u>Question 4</u> : Réaliser les différentes statistiques. Représenter également les boites à moustache des variables Age et IMC ainsi que leur histogramme respectif de telle manière que les deux boites à moustache soient sur la première ligne et les histogrammes sur la seconde ligne. L'histogramme de l'âge doit être de couleur rouge, celui de la variable IMC en bleu. Les axes des abscisses des histogrammes doivent être systématiquement nommés. Les deux histogrammes doivent être avoir un titre. Qu'en concluez-vous ?

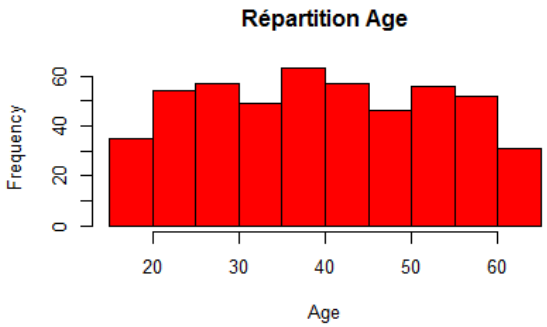
summary(Pression)

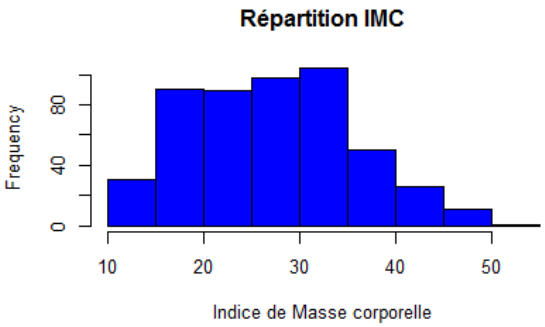
```
 Press_arter    Genre    Fumer         Sport       Age            Alcool          IMC
 1:247          0:264    0:234         0:331    Min.   :18.0    0:327      Min.   :11.00
 2:117          1:236    1:266         1:169    1st Qu.:28.0    1:173      1st Qu.:21.00
 3: 72                                          Median :40.0               Median :27.00
 4: 64                                          Mean   :40.2               Mean   :27.66
                                                3rd Qu.:52.0               3rd Qu.:33.00
                                                Max.   :64.0               Max.   :53.00

 Stress  Sel
 0:326   0:323
 1:174   1:177
```

par(mfrow=c(2,2))
boxplot(Pression$Age)
boxplot(Pression$IMC)
hist(Pression$Age, xlab="Age",main="Répartition Age" ,col="red")
hist(Pression$IMC, xlab="Indice de Masse corporelle",main="Répartition IMC",col="blue")









library(outliers)
grubbs.test(Pression$IMC,type=10, two.sided = TRUE)

```
        Grubbs test for one outlier
data:  Pression$IMC
G = 2.9607, U = 0.9824, p-value = 0.5225
alternative hypothesis: highest value 53 is an outlier
```

La p-value du test est supérieure à 0,05. Par conséquent, la valeur 53 n'est pas considérée comme une valeur atypique au seuil de risque de 5 %. La base de données est conservée en l'état.

Il n'est pas donc nécessaire de refaire les corrélations et les statistiques.
Par contre, on peut vérifier si les variables explicatives qualitatives sont indépendantes entre elles au seuil de risque de 10 %. Idem entre les variables quantitatives et qualitatives.

Indépendance entre les variables qualitatives :

chisq.test(Pression$Genre,Pression$Fumer)

Pearson's Chi-squared test with Yates' continuity correction
data: Pression$Genre and Pression$Fumer
X-squared = 0.82267, df = 1, p-value = 0.3644

- ⇨ Les deux variables (quantitatives) qualitatives Genre et Fumer sont indépendantes au seuil de risque de 10 % (pvalue >0,1)

chisq.test(Pression$Genre,Pression$Sport)
chisq.test(Pression$Genre,Pression$Alcool)
chisq.test(Pression$Genre,Pression$Stress)
chisq.test(Pression$Genre,Pression$Sel)
chisq.test(Pression$Fumer,Pression$Sport)
chisq.test(Pression$Fumer,Pression$Alcool)
chisq.test(Pression$Fumer,Pression$Stress)
chisq.test(Pression$Fumer,Pression$Sel)
chisq.test(Pression$Sport,Pression$Alcool)
chisq.test(Pression$Sport,Pression$Stress)
chisq.test(Pression$Sport,Pression$Sel)
chisq.test(Pression$Alcool,Pression$Stress)
chisq.test(Pression$Alcool,Pression$Sel)
chisq.test(Pression$Stress,Pression$Sel)

- ⇨ Pas de dépendance entre ces différentes variables qualitatives au seuil de risque de 10 %

Lien entre une variable quantitative et une variable qualitative :

t.test(Pression$Age~Pression$Genre)

Welch Two Sample t-test

data: Pression$Age by Pression$Genre
t = -0.11953, df = 493.64, p-value = 0.9049
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.482996 2.198199
sample estimates:
mean in group 0 mean in group 1

40.12879 40.27119

⇨ Pas de différence significative entre l'âge des hommes et des femmes au seuil de risque de 10 % (p = 0,9049)

t.test(Pression$Age~Pression$Fumer)
t.test(Pression$Age~Pression$Sport)
t.test(Pression$Age~Pression$Alcool)
t.test(Pression$Age~Pression$Stress)
t.test(Pression$Age~Pression$Sel)
t.test(Pression$IMC~Pression$Genre)
t.test(Pression$IMC~Pression$Sport)
t.test(Pression$IMC~Pression$Alcool)
t.test(Pression$IMC~Pression$Stress)
t.test(Pression$IMC~Pression$Sel)
t.test(Pression$IMC~Pression$Fumer)

Welch Two Sample t-test

data: Pression$IMC by Pression$Fumer
t = -2.378, df = 490.57, p-value = 0.01779
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-3.3150115 -0.3154107
sample estimates:
mean in group 0 mean in group 1
26.69231 28.50752

⇨ Différence significative d'IMC entre le groupe des personnes fumant et les autres. Attention lors des estimations

<u>Question 5</u> : Quelles sont les variables qui permettent (et ne permettent pas) d'expliquer de manière significative la probabilité que les personnes aient une hypertension artérielle supérieure à la normale?

Pensez à regarder :
- La multicolinéarité entre les variables explicatives utilisées dans l'estimation du modèle
- L'hypothèse de nullité de l'ensemble des coefficients des variables explicatives du modèle
- Les effets marginaux pour les variables explicatives quantitatives
- Les odds-ratios pour les variables explicatives qualitatives
- Le tableau de prédiction et par conséquent le taux d'erreur du modèle estimé
- Le taux de sensibilité et de spécificité du modèle estimé ; le ROC
- La qualité d'ajustement du modèle estimé
- L'existence (ou non) d'observations influençant de manière significative l'estimation
- l'effet de l'IMC en fonction de la pratique du sport sur la probabilité d'avoir une pression artérielle supérieure à la normale
- l'hypothèse d'homoscédasticité des erreurs du modèle estimé

Estimation du modèle

```
modele<-glm(Pression~Age+IMC+Genre+Fumer+Sport+Alcool+Stress+Sel,
data=Pression,family=binomial(link="logit"))
summary(modele)
```

Coefficients:

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*- 1.129702*	*0.465561*	*-2.427*	*0.015244 **
Age	*0.001452*	*0.006972*	*0.208*	*0.835033*
IMC	*0.038262*	*0.011108*	*3.445*	*0.000572 ****
Genre[T.1]	*- 0.262372*	*0.185047*	*-1.418*	*0.156229*
Fumer[T.1]	*0.247853*	*0.186578*	*1.328*	*0.184042*
Sport[T.1]	*- 0.469482*	*0.196272*	*-2.392*	*0.016757 **
Alcool[T.1]	*0.303595*	*0.195542*	*1.553*	*0.120523*
Stress[T.1]	*0.317545*	*0.193947*	*1.637*	*0.101572*
Sel[T.1]	*- 0.070964*	*0.193415*	*-0.367*	*0.713693*

---

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Null deviance: 693.08 on 499 degrees of freedom
Residual deviance: 664.64 on 491 degrees of freedom
AIC: 682.64

➔Les seuls facteurs impactant de manière significative le fait d'avoir une pression artérielle supérieure à la normale est l'IMC (au seuil de risque de 1 %) et le fait de pratiquer du sport de manière intensive (au seuil de risque de 5 %)

Lorsque l'IMC de la personne augmente, la probabilité d'avoir une pression artérielle supérieure à la normale augmente. A l'inverse, lorsque la personne pratique du sport de manière intensive, la probabilité d'avoir une pression artérielle au-dessus de la normale diminue par rapport à une personne qui ne pratique pas le sport de manière intensive.

Multicolinéarité entre les variables explicatives du modèle estimé

```
library(car)
vif(modele)
```

Age	*IMC*	*Genre*	*Fumer*	*Sport*	*Alcool*	*Stress*	*Sel*
1.011917	*1.017545*	*1.007366*	*1.023531*	*1.016458*	*1.017629*	*1.005149*	*1.010955*

➔Pas de multicolinéarité entre les variables explicatives.

Remarque : si on estime le modèle en retirant la variable Fumer (IMC a un effet significatif sur la probabilité d'avoir une pression artérielle supérieure à la normale), les résultats d'estimation changent très peu (cf question 4). On peut donc conserver l'ensemble des variables explicatives

```
modele2<-glm(Pression~Age+IMC+Genre+Sport+Alcool+Stress+Sel, data=Pression,fam
ily=binomial(link="logit"))
summary(modele2)
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.021717   0.457132  -2.235 0.025414 *
Age          0.001392   0.006961   0.200 0.841523
IMC          0.039595   0.011055   3.582 0.000341 ***
Genre1      -0.271152   0.184621  -1.469 0.141915
Sport1      -0.451510   0.195410  -2.311 0.020856 *
Alcool1      0.289393   0.194821   1.485 0.137430
Stress1      0.312822   0.193595   1.616 0.106125
Sel1        -0.088187   0.192779  -0.457 0.647348
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 693.08  on 499  degrees of freedom
Residual deviance: 666.40  on 492  degrees of freedom
AIC: 682.4
```

Intérêt du modèle

chi2<- (modele$null.deviance-modele$deviance)
ddl<-modele$df.null-modele$df.residual
pvalue<-pchisq(chi2,ddl,lower.tail=F)
print(pvalue)

[1] 0.0003974904

➔Le modèle a un intérêt car p<0,05

Calcul des odd-ratios

exp(coef(modele))

(Intercept)	*Age*	*IMC*	*Genre[T.1]*	*Fumer[T.1]*	*Sport[T.1]*
0.3231297	*1.0014531*	*1.0390039*	*0.7692245*	*1.2812715*	*0.6253263*

Alcool[T.1]	*Stress[T.1]*	*Sel[T.1]*
1.3547200	*1.3737510*	*0.9314953*

➔La personne pratiquant le sport (de manière intensive) a 1,6 fois (1/0,625) moins de chance d’avoir une pression artérielle supérieure à la normale par rapport à une personne ne pratiquant pas le sport (de manière intensive).

Calcul des effets marginaux

mean(dlogis(predict(modele,type="link")))*coef(modele)

(Intercept)	*Age*	*IMC*	*Genre[T.1]*	*Fumer[T.1]*
-0.2667232324	*0.0003428204*	*0.0090337932*	*-0.0619462863*	*0.0585182296*

Sport[T.1]	*Alcool[T.1]*	*Stress[T.1]*	*Sel[T.1]*
-0.1108449355	*0.0716789128*	*0.0749725694*	*-0.0167546786*

➔L'augmentation d'une unité de l'IMC fera augmenter de 0,00903 la probabilité d'avoir une pression artérielle anormale.

Tableau de prévision et % d'erreur du modèle estimé

```
pred.proba<-predict(modele,type="response")
pred.moda<-factor(ifelse(pred.proba>0.5,"1","0"))
mc<-table(Pression$Pression,pred.moda)
print(mc)
```

```
  pred.moda
     0   1
0  143 104
1   99 154
```

```
err<-(mc[2,1]+mc[1,2])/sum(mc)
print(err)
```

```
[1] 0.406
```

➔Le taux d'erreur du modèle est de 40,6 % , ce qui est élevé.

Calcul du taux de sensibilité et de spécificité pour le modèle estimé

```
Sensibilite<-mc[2,2]/(mc[2,1]+mc[2,2])
print(Sensibilite)
[1] 0.6086957

Specificite<-mc[1,1]/(mc[1,1]+mc[1,2])
print(Specificite)
[1] 0.5789474
```

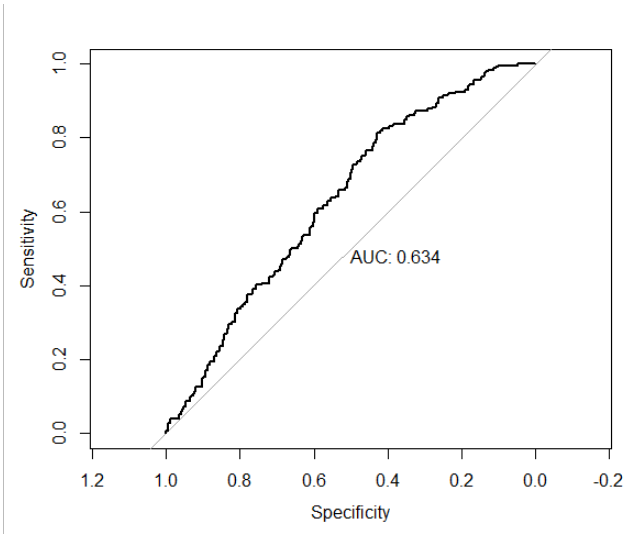
Il existe également la fonction suivante :

```
library(pscl)
hitmiss(modele)
```

```
Classification Threshold = 0.5
          y=0 y=1
yhat=0   143  99
yhat=1   104 154

Percent Correctly Predicted = 59.4%
Percent Correctly Predicted = 57.89%, for y = 0
Percent Correctly Predicted = 60.87%  for y = 1

Null Model Correctly Predicts 50.6%

[1] 59.40000   57.89474  60.86957
```

⇨ Le modèle prédit correctement les valeurs dans 59,4 % des cas. Il prédit de manière assez similaire y=1 (60,9%) et y=0 (57,9%).

library(pROC)
pred <-predict(modele)
Test_roc=roc(Pression$Pression ~pred, plot=TRUE, print.auc=TRUE)



Qualité du modèle estimé

R2_Mc_Fadden<-1-(modele$deviance/modele$null.deviance)
R2_Mc_Fadden

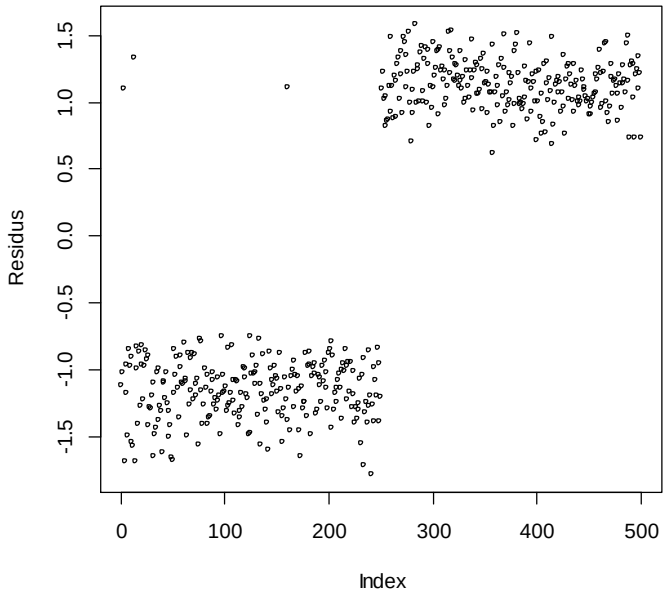
[1] 0.04103393

➔La qualité d'ajustement du modèle est très faible.

Observations influençant (ou non) de manière significative l'estimation

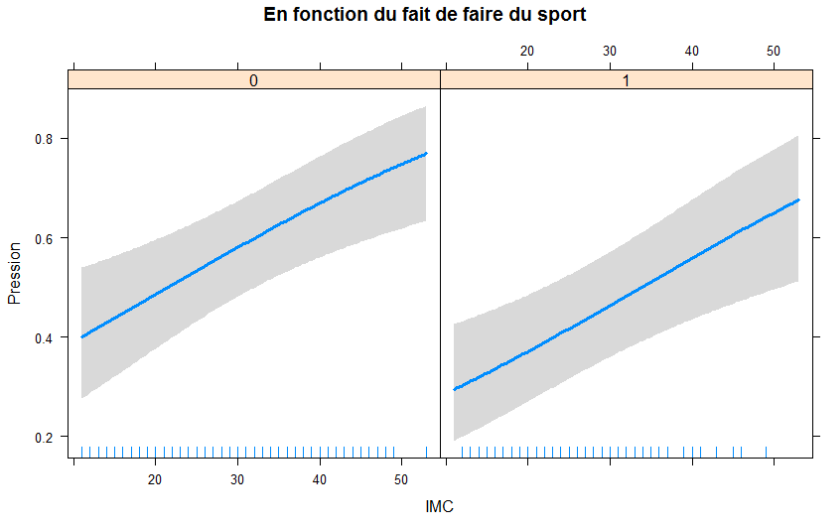
plot(rstudent(modele),type="p",cex=0.5,ylab="Residus")
abline(h=c(-2,2))

Dans ce modèle, les résidus de l'estimation sont compris entre -2 et 2. Il n'y a donc pas d'observations influençant de manière significative l'estimation réalisée.



Effet de l'IMC en fonction de la pratique du sport sur la probabilité d'avoir une pression artérielle supérieure à la normale

```
library(visreg)
visreg(modele, "IMC", by="Sport",scale="response", main="En fonction du fait de faire du sport")
```



Vérification de l'hypothèse d'homoscédasticité des erreurs du modèle estimé

```
modeleh <- hetglm(Pression~Age+IMC+Genre+Fumer+Sport+Alcool+Stress+Sel|Age+IMC+Genre+Fumer+Sport+Alcool+Stress+Sel,data=Pression,family=binomial(logit))
summary(modeleh)
```

```
Coefficients (binomial model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.8892659  0.8644535  -1.029    0.304
Age         -0.0007726  0.0058623  -0.132    0.895
IMC          0.0354760  0.0358665   0.989    0.323
```

```
Genre1      -0.5406332  0.6990728  -0.773    0.439
Fumer1       0.4016938  0.4005368   1.003    0.316
Sport1      -0.4732801  0.4884742  -0.969    0.333
Alcool1      0.6682805  0.6767303   0.988    0.323
Stress1      0.3772290  0.3987452   0.946    0.344
Sel1        -0.0802486  0.1487222  -0.540    0.589

Latent scale model coefficients (with log link):
        Estimate Std. Error z value Pr(>|z|)
Age     -0.04267    0.01538  -2.774  0.00553 **
IMC      0.04555    0.02227   2.045  0.04084 *
Genre1   1.21577    0.63877   1.903  0.05700 .
Fumer1   0.42746    0.37010   1.155  0.24809
Sport1  -0.92888    0.36593  -2.538  0.01114 *
Alcool1  0.86264    0.46200   1.867  0.06188 .
Stress1  0.92687    0.45910   2.019  0.04350 *
Sel1    -0.54333    0.35315  -1.539  0.12392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -317.6 on 17 Df
LR test for homoskedasticity: 29.45 on 8 Df, p-value: 0.0002638
```

vif(modeleh)

```
  Age    IMC      Genre     Fumer     Sport    Alcool    Stress       Sel
1.207 25.158     10.723    12.553   17.920   20.062     6.523     1.715
```

- ⇨ Il existe un problème de multicolinéarité. On peut réestimer le modèle en supprimant les variables Sel, Fumer comme variables expliquant l'hétéroscédasticité (car p-value du LR Test (0,0002638 <0,05))

modeleh<hetglm(Pression~Age+Fumer+IMC+Genre+Sport+Alcool+Stress+Sel|Age+IMC+Genre+Sport+Alcool+Stress, data=Pression,family=binomial(logit))
summary(modeleh)

```
              Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.9396078  0.8421138  -1.116    0.265
Age         -0.0005517  0.0052637  -0.105    0.917
Fumer1       0.3637283  0.3391551   1.072    0.284
IMC          0.0358058  0.0330597   1.083    0.279
Genre1      -0.3806273  0.4177923  -0.911    0.362
Sport1      -0.4138311  0.4008273  -1.032    0.302
Alcool1      0.5585600  0.5449888   1.025    0.305
Stress1      0.3857230  0.3752305   1.028    0.304
Sel1        -0.0777501  0.1386021  -0.561    0.575

Latent scale model coefficients (with log link):
        Estimate Std. Error z value Pr(>|z|)
Age     -0.03420    0.01435  -2.383  0.01719 *
IMC      0.03577    0.02130   1.679  0.09308 .
Genre1   0.81073    0.47344   1.712  0.08682 .
Sport1  -1.04745    0.36258  -2.889  0.00387 **
Alcool1  0.94463    0.49050   1.926  0.05412 .
Stress1  0.64513    0.42436   1.520  0.12845
Log-likelihood:  -319 on 15 Df
LR test for homoskedasticity: 26.65 on 6 Df, p-value: 0.0001681
```

vif(modeleh)

Age	*Fumer*	*IMC*	*Genre*	*Sport*	*Alcool*	*Stress*	*Sel*
1.179726	*8.285029*	*21.020138*	*6.584095*	*10.361202*	*9.894046*	*6.929730*	*1.365860*

⇨ Il existe toujours un problème de multicolinéarité. On peut supprimer la variable Stress comme variable explicative dans l'analyse de la variance

modeleh<-hetglm(Pression~Age+Fumer+IMC+Genre+Sport+Alcool+Stress+Sel| Age+IMC+Genre+Sport+Alcool,data=Pression,family=binomial(logit))
summary(modeleh)

```
Coefficients (binomial model with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.240811   1.109578  -1.118    0.263
Age          0.004307   0.006124   0.703    0.482
Fumer1       0.521731   0.496889   1.050    0.294
IMC          0.034448   0.033205   1.037    0.300
Genre1      -0.459142   0.524957  -0.875    0.382
Sport1      -0.465484   0.474262  -0.981    0.326
Alcool1      0.702858   0.721571   0.974    0.330
Stress1      0.465301   0.447327   1.040    0.298
Sel1        -0.082165   0.161564  -0.509    0.611

Latent scale model coefficients (with log link):
        Estimate Std. Error z value Pr(>|z|)
Age     -0.03807    0.01548  -2.459  0.01392 *
IMC      0.05818    0.02382   2.443  0.01459 *
Genre1   0.88142    0.48009   1.836  0.06637 .
Sport1  -1.14449    0.38601  -2.965  0.00303 **
Alcool1  1.21100    0.55626   2.177  0.02948 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -319.8 on 14 Df
LR test for homoskedasticity: 24.97 on 5 Df, p-value: 0.0001411
```

vif(modeleh)

Age	Fumer	IMC	Genre	Sport	Alcool	Stress	Sel
1.241	13.951	18.678	6.967	10.557	8.917	13.435	1.492

⇨ Il existe toujours un problème de multicolinéarité. On peut supprimer la variable Genre comme variable explicative dans l'analyse de la variance afin de raisonner au seuil de risque de 5%

modeleh <- hetglm(Pression~Age+Fumer+IMC+Genre+Sport+Alcool+Stress+Sel| Age+IMC+Sport+Alcool,data=Pression,family=binomial(logit))
summary(modeleh)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.554460   0.526219  -1.054    0.292
Age         -0.002264   0.003891  -0.582    0.561
Fumer1       0.172824   0.170443   1.014    0.311
IMC          0.024684   0.022817   1.082    0.279
Genre1      -0.119765   0.134051  -0.893    0.372
Sport1      -0.233911   0.232429  -1.006    0.314
Alcool1      0.296526   0.322905   0.918    0.358
Stress1      0.166051   0.168038   0.988    0.323
Sel1        -0.057737   0.093758  -0.616    0.538
```

```
Latent scale model coefficients (with log link):
        Estimate Std. Error z value Pr(>|z|)
Age     -0.01102    0.01324  -0.832  0.40524
IMC      0.00253    0.02114   0.120  0.90471
Sport1  -1.28960    0.41296  -3.123  0.00179 **
Alcool1  1.15257    0.63165   1.825  0.06804 .

Log-likelihood: -321.4 on 13 Df
LR test for homoskedasticity: 21.83 on 4 Df, p-value: 0.0002171
```

```
vif(modeleh)
   Age     Fumer       IMC     Genre     Sport    Alcool    Stress       Sel
 1.617     5.049    20.962     3.212     5.394     4.329     4.838     1.480
```

⇨ Il existe toujours un problème de multicolinéarité. On peut supprimer les variables Age et IMC comme variables explicatives dans l'analyse de la variance

modeleh <- hetglm(Pression~Age+Fumer+IMC+Genre+Sport+Alcool+Stress+Sel|Sport+Alcool,data=Pression,family=binomial(logit))
summary(modeleh)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.735566   0.421886  -1.744   0.0812 .
Age         -0.003889   0.004330  -0.898   0.3692
Fumer1       0.229763   0.129826   1.770   0.0768 .
IMC          0.034612   0.013452   2.573   0.0101 *
Genre1      -0.140055   0.113893  -1.230   0.2188
Sport1      -0.324114   0.158667  -2.043   0.0411 *
Alcool1      0.360598   0.292958   1.231   0.2184
Stress1      0.199912   0.123119   1.624   0.1044
Sel1        -0.079559   0.108159  -0.736   0.4620

Latent scale model coefficients (with log link):
        Estimate   Std. Error z value Pr(>|z|)
Sport1   -1.4204     0.4256  -3.337 0.000846 ***
Alcool1   1.2853     0.6700   1.918 0.055053 .

Log-likelihood: -321.7 on 11 Df
LR test for homoskedasticity: 21.33 on 2 Df, p-value: 2.334e-05
```

vif(modeleh)

```
   Age     Fumer       IMC     Genre     Sport    Alcool    Stress       Sel
1.135235 1.614468 3.824687 1.285298 1.170770 1.598142 1.442269 1.082060
```

⇨ Il n'existe plus de problème de multicolinéarité. On peut supprimer néanmoins la variable Alcool si on veut raisonner au seuil de risque de 5%

modeleh <- hetglm(Pression~Age+Fumer+IMC+Genre+Sport+Alcool+Stress+Sel|Sport,data=Pression,family=binomial(logit))
summary(modeleh)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4801335  0.3697616  -1.298   0.1941
Age         -0.0002117  0.0026030  -0.081   0.9352
Fumer1       0.1252542  0.0870329   1.439   0.1501
IMC          0.0195724  0.0099976   1.958   0.0503 .
Genre1      -0.1072459  0.0823926  -1.302   0.1930
Sport1      -0.2366873  0.1196217  -1.979   0.0479 *
Alcool1      0.1751612  0.1062233   1.649   0.0991 .
Stress1      0.1493869  0.0969265   1.541   0.1233
Sel1         0.0040835  0.0674768   0.061   0.9517
```

```
Latent scale model coefficients (with log link):
        Estimate Std. Error z value Pr(>|z|)
Sport1  -1.6483     0.5352   -3.08  0.00207 **

Log-likelihood: -324.9 on 10 Df
LR test for homoskedasticity: 14.92 on 1 Df, p-value: 0.0001125
```

```
vif(modeleh)
     Age    Fumer      IMC    Genre    Sport   Alcool   Stress      Sel
1.026025 1.815510 5.501770 1.647075 1.064873 2.520613 2.060346 1.022056
```

Remarque : si on raisonne dès le début au seuil de risque de 5% / variance et par élimination, on obtient également le modèle ci-dessus.

La variable Sport a un impact significatif au seuil de risque de 5 % sur la probabilité d'avoir une tension artérielle supérieure à la normale. Le fait de faire du sport fait diminuer cette probabilité
La variable IMC et Alcool ont un impact significatif positif quant à elles au seuil de risque de 10% sur cette probabilité.

L'hypothèse d'homoscédasticité des erreurs doit être refusée au seuil de risque de 1% (et donc de 5%) car p-value =0,0001125

Qualité du modèle

h1c <- hetglm(Pression~1,data=Pression,family=binomial(logit))
(R2McFAdden<-1-(modeleh$loglik/h1c$loglik))

[1] 0.06255416
(meilleure qualité d'ajustement que dans le modèle où les erreurs sont supposées homoscédastiques : 0,04)

Estimation du modèle à erreurs supposées homoscédastiques et comparaison avec le modèle prenant en compte l'hétéroscédasticité des erreurs

h1h<- hetglm(Pression~Age+IMC+Genre+Fumer+Sport+Alcool+Stress+Sel|
1,data=Pression,family=binomial(logit))
library(lmtest)
lrtest(h1h,modeleh)

Likelihood ratio test

Model 1: Pression ~ Age + IMC + Genre + Fumer + Sport + Alcool + Stress + Sel | 1
Model 2: Pression ~ Age + Fumer + IMC + Genre + Sport + Alcool + Stress + Sel | Sport

```
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -332.32
2  10 -324.86  1 14.915  0.0001125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

⇨ Au seuil de risque de 1% (et donc de 5%), le modèle modeleh (prise en compte de l'hé téroscédasticité des erreurs) est préféré au modèle h1h (modèle où les erreurs sont supp osées homoscédastiques). Il faut donc conserver le modèle hétéroscédastique et interpr éter ses résultats