Correction TD2

Econométrie des variables qualitatives 1
Modèles multinomiaux

Exercice : Le choix des marques et ses déterminants

Vous disposez pour 735 personnes de leur choix en termes de marque (Marque1, Marque2, Marque3) ainsi que de leurs caractéristiques et du prix de ces 3 marques. Les données sont disponibles dans le fichier Marque.xls (sous Madoc).

Choice : choix de la personne en termes de marque (Marque1, Marque2, Marque3)
Prix.Marque1 : prix de la marque 1 (en €)
Prix.Marque2 : prix de la marque 2 (en €)
Prix.Marque3 : prix de la marque 3 (en €)
Femme = 1 si la personne interrogée est une femme, 0 sinon
Age : Age de personne interrogée (en année)

Question 1 : Sous quel type de format est enregistrée la base de données? Importer la base sous le logiciel R. Nommer la base : Marque

La base de données est enregistrée sous format court car il y a 735 lignes. Il existe 3 colonnes pour la variable Prix.

```
getwd()
setwd("C:/Users/travers-m/Desktop/Cours_2020_2021/Econometrie_variables_qualitatives_M1_EKAP_M2_CODEME/TD/Bases")
library(readxl)

Marque <-read_excel("Marque.xls",sheet="Feuil1",col_names=TRUE)

str(Marque)
```

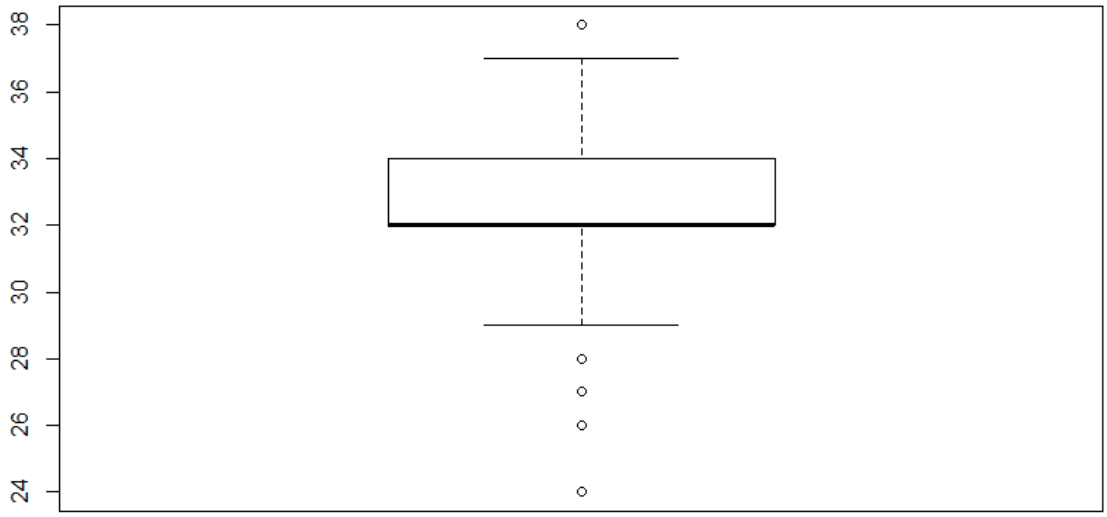
```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':	735 obs. of  6 variables:
 $ Choice      : chr  "Marque1" "Marque1" "Marque1" "Marque1" ...
 $ Femme       : Factor w/ 2 levels "0","1": 1 1 1 2 2 2 1 1 2 1 ...
 $ Age         : num  24 26 26 27 27 27 27 27 27 27 ...
 $ Prix.Marque1: num  5 6 7 7 5 7 8 7 6 7 ...
 $ Prix.Marque2: num  6 3 4 4 4 4 4 4 2 6 ...
 $ Prix.Marque3: num  1 4 3 1 2 5 1 4 4 3 ...
```

Si problème de lecture de la base :

```
Marque <- as.data.frame(Marque)
```

<u>Question 2</u> : Vérifier s'il n'existe pas de valeurs atypiques pour la variable Age

boxplot(Marque$Age)



➔Il existe des valeurs potentiellement atypiques. Il faut donc vérifier cela via le test de Rosner.

```
y = Marque$Age

rval = function(y){
  ares = abs(y - mean(y))/sd(y)
  df = data.frame(y, ares)
  r = max(df$ares)
  list(r, df)}
n = length(y)
alpha = 0.05
lam = c(1:10)
R = c(1:10)
for (i in 1:10){
  if(i==1){
    rt = rval(y)
    R[i] = unlist(rt[1])
    df = data.frame(rt[2])
    newdf = df[df$ares!=max(df$ares),]}
  else if(i!=1){
    rt = rval(newdf$y)
    R[i] = unlist(rt[1])
    df = data.frame(rt[2])
    newdf = df[df$ares!=max(df$ares),]}
  p = 1 - alpha/(2*(n-i+1))
  t = qt(p,(n-i-1))
  lam[i] = t*(n-i) / sqrt((n-i-1+t**2)*(n-i+1))
}
newdf = data.frame(c(1:10),R,lam)
names(newdf)=c("No. Outliers","Test Stat.", "Critical Val.")
newdf
```

No.	*Outliers*	*Test Stat.*	*Critical Val.*
1	*1*	*3.814803*	*3.962965*
2	*2*	*2.990624*	*3.962619*
3	*3*	*2.594613*	*3.962273*
4	*4*	*2.273454*	*3.961926*
5	*5*	*2.328447*	*3.961579*
6	*6*	*2.224474*	*3.961231*
7	*7*	*2.176756*	*3.960883*
8	*8*	*1.947232*	*3.960534*
9	*9*	*2.358979*	*3.960185*
10	*10*	*2.407364*	*3.959835*

Au seuil de risque de 5%, il n'existe pas de valeurs atypiques. On peut donc conserver la base initiale.

<u>Question 3</u> : Transformer la variable de type qualitatif en facteur sous R.

Marque$Femme<-as.factor(Marque$Femme)

<u>Question 4</u> : Quelles sont la ou les variables liée(s) au choix? Quelles sont la ou les variables liée(s) aux individus ?

La variable explicative liée au choix est la variable Prix, puisque le prix varie selon la marque. Les variables explicatives liées aux individus sont les variables : Femme et Age qui ne dépendent pas du choix (de la Marque) mais de l'individu.

<u>Question 5</u> : Comment appelle-t-on un modèle de choix où on ne prend que les variables liées aux individus ? Faire l'estimation d'un tel modèle avec la fonction mlogit et vglm et interpréter les résultats obtenus. Au préalable, pour la fonction mlogit, transformer la base Marque dans le format adéquat (nom de la nouvelle base : Marqueb). Combien de lignes doit avoir cette nouvelle base ? Quel choix est-pris par défaut comme référence pour l'interprétation des résultats ?

On parle de modèle multinomial simple lorsque les variables explicatives sont uniquement des variables fonction de l'individu. On peut estimer ce type de modèle via la fonction mlogit et vglm. Pour utiliser la fonction mlogit, il faut au préalable modifier le format de la base Marque. Cette nouvelle base doit avoir 735 (individus)*3(Marques) = 2205 lignes

library(mlogit)
Marqueb<-mlogit.data(Marque, shape="wide",varying=3:5,choice="Choice")
head(Marqueb,8)

	Choice	*Femme*	*Age*	*alt*	*Prix*	*chid*
1	*TRUE*	*0*	*24*	*Marque1*	*5*	*1*
2	*FALSE*	*0*	*24*	*Marque2*	*6*	*1*
3	*FALSE*	*0*	*24*	*Marque3*	*1*	*1*
4	*TRUE*	*0*	*26*	*Marque1*	*6*	*2*
5	*FALSE*	*0*	*26*	*Marque2*	*3*	*2*
6	*FALSE*	*0*	*26*	*Marque3*	*4*	*2*
7	*TRUE*	*0*	*26*	*Marque1*	*7*	*3*
8	*FALSE*	*0*	*26*	*Marque2*	*4*	*3*

Estimation du modèle multinomial simple via la fonction mlogit

ml.Marquems<-mlogit(Choice~0|Age+Femme,data=Marqueb)
summary(ml.Marquems)

Frequencies of alternatives:
Marque1 Marque2 Marque3
0.28163 0.41769 0.30068

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
Marque2:(intercept)	-11.774478	1.774612	-6.6350	3.246e-11 ***
Marque3:(intercept)	-22.721201	2.058028	-11.0403	< 2.2e-16 ***
Marque2:Age	0.368201	0.055003	6.6942	2.169e-11 ***
Marque3:Age	0.685902	0.062627	10.9523	< 2.2e-16 ***
Marque2:Femme1	0.523813	0.194247	2.6966	0.007004 **
Marque3:Femme1	0.465939	0.226090	2.0609	0.039316 *

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

Log-Likelihood: -702.97
McFadden R^2: 0.11676
Likelihood ratio test : chisq = 185.85 (p.value = < 2.22e-16)

➔ La marque1 a été automatiquement mise comme choix de référence.

Estimation du modèle multinomial simple via la fonction vglm

Cette estimation se fait à partir de la base initiale Marque

library(VGAM)
Fit<-vglm(Choice~Age+Femme,multinomial(refLevel=1),data=Marque)
summary(Fit)

Pearson residuals:

	Min	1Q	Median	3Q	Max
log(mu[,2]/mu[,1])	-5.152	-0.6183	-0.4108	0.9800	1.779
log(mu[,3]/mu[,1])	-5.836	-0.6697	-0.2441	0.6138	7.287

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-11.77466	1.77460	-6.635	3.24e-11 ***
(Intercept):2	-22.72140	2.05802	-11.040	< 2e-16 ***
Age:1	0.36821	0.05500	6.694	2.17e-11 ***
Age:2	0.68591	0.06263	10.952	< 2e-16 ***
Femme1:1	0.52381	0.19425	2.697	0.0070 **
Femme1:2	0.46594	0.22609	2.061	0.0393 *

Number of linear predictors: 2
Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])

Residual deviance: 1405.941 on 1464 degrees of freedom
Log-likelihood: -702.9707 on 1464 degrees of freedom
Number of iterations: 5

Interprétation

Si l'âge de l'individu augmente d'une année, la probabilité de choisir la marque 2 augmente par rapport à la marque 1. Il en est de même pour la marque 3.
Ces deux effets sont significatifs au seuil de risque de 1%.

Le fait d'être une femme augmente la probabilité de choisir la marque 2 par rapport à la marque 1. Il en est de même pour la marque 3.
Le premier effet est significatif au seuil de risque de 1%. Le second effet est significatif au seuil de risque de 5%.

La qualité d'ajustement du modèle est de 0,117, ce qui est correct pour ce type de modèle. Il y a un intérêt à estimer ce modèle (Likelihood ratio test : p.value = < 2.22e-16 < 0,05)

Question 6 : Estimer le modèle avec l'ensemble des variables explicatives en utilisant la fonction mlogit. Comment appelle-t-on ce type de modèle ? Le fait d'avoir rajouter la variable Prix comme variable explicative est-il pertinent ?

On parle de modèle multinomial général puisque les variables dépendent à la fois des individus (Femme et Age) et du choix (Prix).

ml.Marquemg<-mlogit(Choice~Prix|Age+Femme,data=Marqueb)
summary(ml.Marquemg)

Frequencies of alternatives:
Marque1 Marque2 Marque3
0.28163 0.41769 0.30068

Coefficients :

	Estimate	*Std. Error*	*t-value*	*Pr(>|t|)*
Marque2:(intercept)	*-11.728204*	*1.775242*	*-6.6065*	*3.934e-11 ****
Marque3:(intercept)	*-22.632040*	*2.062747*	*-10.9718*	*< 2.2e-16 ****
Prix	*0.018270*	*0.032673*	*0.5592*	*0.57604*
Marque2:Age	*0.368732*	*0.054981*	*6.7066*	*1.992e-11 ****
Marque3:Age	*0.685674*	*0.062602*	*10.9530*	*< 2.2e-16 ****
Marque2:Femme1	*0.524654*	*0.194268*	*2.7007*	*0.00692 ***
Marque3:Femme1	*0.466038*	*0.226105*	*2.0612*	*0.03929 **

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

Log-Likelihood: -702.81
McFadden R^2: 0.11695
Likelihood ratio test : chisq = 186.16 (p.value = < 2.22e-16)

➔La variable Prix n'a pas d'impact significatif au seuil de risque de 10 % sur le choix de la marque

library(lmtest)
lrtest(ml.Marquems,ml.Marquemg)

Likelihood ratio test

Model 1: Choice ~ 0 | Age + Femme
Model 2: Choice ~ Prix | Age + Femme

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	6	-702.97			
2	7	-702.81	1	0.3127	0.576

➔Ce test comparant le modèle multinomial simple et général indique que le modèle multinomial n'améliore pas la qualité de l'estimation au seuil de risque de 5%.
Par conséquent, pour la suite des questions, le modèle multinomial simple sera utilisé, ce qui permet ainsi de pouvoir utiliser la fonction vglm.

<u>Question 7</u> : En prenant le modèle retenu à la question 5 (via la fonction vglm), calculer les odds-ratios des variables explicatives (via la fonction vglm). Interpréter les résultats

exp(coef(Fit))

(Intercept):1	(Intercept):2	Age:1	Age:2	Femme:1	Femme:2
7.697192e-06	1.355886e-10	1.445140	1.985574	1.688456	1.593514

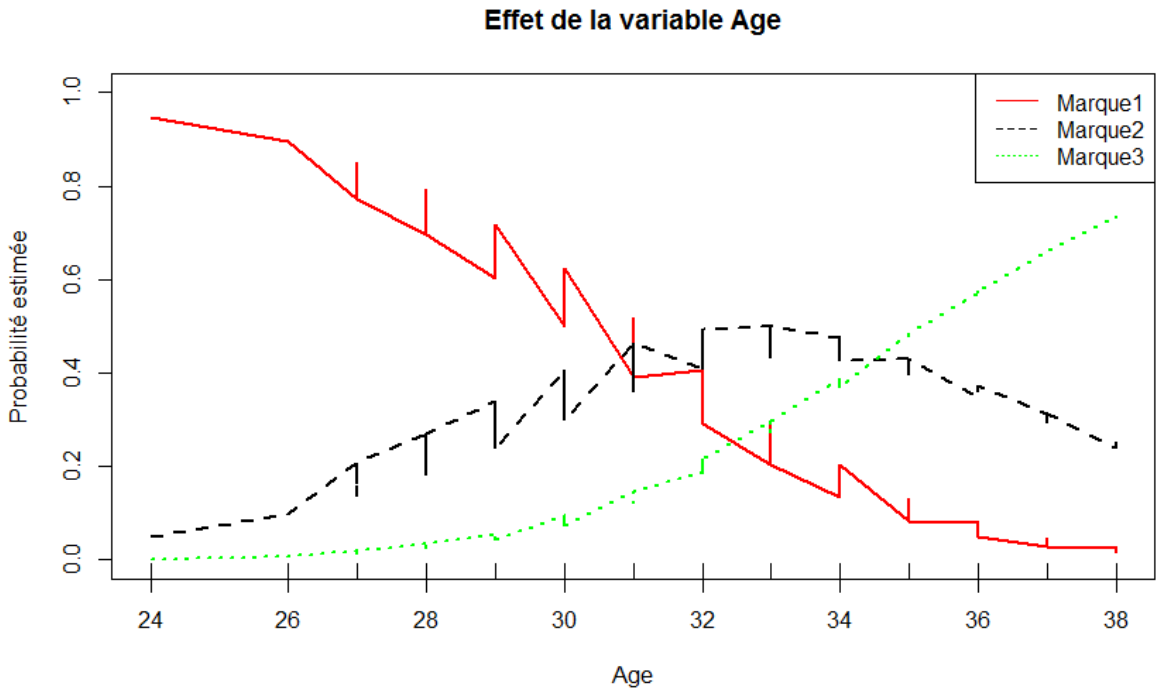
<u>Interprétation :</u>

Un individu dont l'âge augmente d'un an a 1,45 fois plus de chance de choisir la marque 2 (/ Marque1) par rapport à celui dont l'âge ne change pas.
Un individu dont l'âge augmente d'un an a 1,99 fois plus de chance de choisir la marque 3 (/ Marque1) par rapport à celui dont l'âge ne change pas.

Une femme a 1,69 fois plus de chance de choisir la marque 2 (/ Marque1) qu'un homme.
Une femme a 1,59 fois plus de chance de choisir la marque 3 (/ Marque1) qu'un homme.

<u>Question 8</u> : En prenant le modèle retenu à la question 5, représenter graphiquement les effets de la variable Age sur la probabilité d'avoir les différentes marques (via la fonction vglm)

```
mycol <- c("red", "green")
ooo <- with(Marque, order(Age))
with(Marque, matplot(Age[ooo], fitted(Fit)[ooo,], ylim = c(0,1),
                     xlab = "Age", ylab = "Probabilité estimée",
                     main = " Effet de la variable Age ", type = "l", lwd = 2, col = c(mycol[1],
"black", mycol[-1])))
with(Marque, rug(Age))
legend("topright", col = c(mycol[1], "black", mycol[-1]), lty=1:3
,legend=colnames(Fit@y))
```



<u>Question 9</u> : Estimer le modèle de la question 5 (via la fonction mlogit) en tenant compte de l'hétéroscédasticité des erreurs. Doit-on conserver ce modèle ? Comparer les résultats de ce modèle au modèle où les erreurs sont supposées homoscédastiques.

ml.MarqueHet<-mlogit(Choice~0|Age+Femme,heterosc = TRUE,data=Marqueb)
summary(ml.MarqueHet)

Frequencies of alternatives:
Marque1 Marque2 Marque3
0.28163 0.41769 0.30068

Coefficients :

	Estimate	*Std. Error*	*t-value*	*Pr(>|t|)*
Marque2:(intercept)	*-10.998914*	*1.389508*	*-7.9157*	*2.442e-15 ****
Marque3:(intercept)	*-23.920272*	*9.460678*	*-2.5284*	*0.011459 **
Marque2:Age	*0.354164*	*0.042337*	*8.3653*	*< 2.2e-16 ****
Marque3:Age	*0.712822*	*0.257591*	*2.7673*	*0.005653 ***
Marque2:Femme	*0.448271*	*0.151836*	*2.9523*	*0.003154 ***
Marque3:Femme	*0.518847*	*0.251373*	*2.0641*	*0.039012 **
sp.Marque2	*0.304102*	*0.332928*	*0.9134*	*0.361023*
sp.Marque3	*1.222708*	*0.790282*	*1.5472*	*0.121820*

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

Log-Likelihood: -699.75
McFadden R^2: 0.1208
Likelihood ratio test : chisq = 192.28 (p.value = < 2.22e-16)

lrtest(ml.Marquems,ml.MarqueHet)
Likelihood ratio test

Model 1: Choice ~ 0 | Age + Femme
Model 2: Choice ~ 0 | Age + Femme

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	6	-702.97			
2	8	-699.75	2	6.4325	0.0401 *

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

➔Au seuil de risque de 5%, le modèle multinomial simple avec prise en compte de l'hétéroscédasticité doit être conservé.

Rappel : modèle multinomial simple avec erreurs homoscédastiques

Coefficients :

	Estimate	Std. Error	t-value	Pr(>|t|)
Marque2:(intercept)	-11.774478	1.774612	-6.6350	3.246e-11 ***
Marque3:(intercept)	-22.721201	2.058028	-11.0403	< 2.2e-16 ***
Marque2:Age	0.368201	0.055003	6.6942	2.169e-11 ***
Marque3:Age	0.685902	0.062627	10.9523	< 2.2e-16 ***
Marque2:Femme1	0.523813	0.194247	2.6966	0.007004 **
Marque3:Femme1	0.465939	0.226090	2.0609	0.039316 *

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

Log-Likelihood: -702.97
McFadden R^2: 0.11676
Likelihood ratio test : chisq = 185.85 (p.value = < 2.22e-16)

Comparaison des résultats du modèle multinomial simple avec prise en compte de l'hétéroscédasticité des erreurs / modèle multinomial simple :

En termes de significativité, il n'y a pas de différence pour les deux variables Age et Femme. Les coefficients des variables explicatives significatives ne changent pas de signe et varient peu.
La qualité d'ajustement du modèle est légèrement supérieure (0,121) à celle du modèle avec erreurs homoscédastiques (0,117).

<u>Question 10</u> : Supposons maintenant que le choix de la personne se fasse entre la marque 2 et un ensemble de marques constitué par les marques 1 et 3. Pour ce faire, réintroduire le prix comme variable explicative et prendre la marque 2 comme référence. L'hypothèse IIA est-elle vérifiée ?

ml.Marque<-mlogit(Choice~Prix|Age+Femme,data=Marqueb,reflevel="Marque2")
scoretest(ml.Marque,
nests=list(type1="Marque2",type2=c("Marque3","Marque1")),unscaled=TRUE)

```
        score test

data:  type1, type2
chisq = 10.701, df = 2, p-value = 0.004746
alternative hypothesis: nested model
```

nl.Marque<-
mlogit(Choice~Prix|Age+Femme,data=Marqueb,reflevel="Marque2",nests=list(Type1="Marque2",Type2=c("Marque3","Marque1")),unscaled=TRUE)
lrtest(nl.Marque)

```
Likelihood ratio test

Model 1: Choice ~ Prix | Age + Femme
Model 2: Choice ~ Prix | Age + Femme
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -697.79
2   7 -702.81 -2 10.045   0.006588 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

summary(nl.Marque)

```
Call:
mlogit(formula = Choice ~ Prix | Age + Femme, data = Marqueb,
    reflevel = "Marque2", nests = list(Type1 = "Marque2",
        Type2 = c("Marque3", "Marque1")), unscaled = TRUE)

Frequencies of alternatives:
Marque2 Marque1 Marque3
0.41769 0.28163 0.30068

bfgs method
15 iterations, 0h:0m:0s
g'(-H)^-1g = 3.74E-07
gradient close to zero

Coefficients :
                   Estimate Std. Error z-value  Pr(>|z|)
Marque1:(intercept)  9.945089   1.343836  7.4005 1.357e-13 ***
Marque3:(intercept) -9.492209   1.081495 -8.7769 < 2.2e-16 ***
Prix                 0.030749   0.025286  1.2160  0.223968
Marque1:Age         -0.327465   0.041553 -7.8806 3.331e-15 ***
Marque3:Age          0.262768   0.031572  8.3228 < 2.2e-16 ***
Marque1:Femme       -0.353932   0.125621 -2.8174  0.004841 **
Marque3:Femme        0.089800   0.120701  0.7440  0.456888
iv:Type1            -0.737600   1.923131 -0.3835  0.701319
iv:Type2             2.225533   0.530735  4.1933 2.749e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Log-Likelihood: -697.79
McFadden R^2:  0.12326
Likelihood ratio test : chisq = 196.21 (p.value = < 2.22e-16)
```

Interprétation :

Le fait d'être une femme n'a plus d'impact significatif au seuil de risque de 10% sur le fait de choisir marque 3 par rapport à la marque 2.