Chapitre 2 :
Modèle multinomial ordonné

1) Présentation théorique

➔ Modèle dans lequel la variable à expliquer prend plusieurs modalités ayant un ordre naturel.

➔ Les modalités doivent être identiques pour tous les individus et être mutuellement exclusives

➔ Généralisation du modèle dichotomique avec une variable Y prenant j modalités allant de j= 1 à k pour chaque individu i=1, …, n

Soient p variables explicatives.

Il existe alors plusieurs seuils $\alpha_1,\ldots, \alpha_{k-1}$ tels que :

$$Y_i = \begin{cases} 1 \text{ si } Y_i^* < \alpha_1 \\ 2 \text{ si } \alpha_1 \le Y_i^* < \alpha_2 \\ \ldots \\ k \text{ si } Y_i^* \ge \alpha_{k-1} \end{cases}$$

Où : $\alpha_{j+1} > \alpha_j$

la variable latente Y^*_i dépend linéairement de variables explicatives quantitatives ou qualitatives.

$$Y_i^* = \sum_{m=i}^{p} \beta_m x_{im} + u_i = X_i\beta + \varepsilon_i \qquad i = 1,...,n$$

➔ Les probabilités des différentes modalités j dépendent des coefficients estimés des variables explicatives et des constantes associées à chaque niveau de Y (valeur seuil).

➔ Soit Y ayant 3 valeurs ordonnées (1< 2< 3) :

$$\text{Prob}(Y_i = 1 \mid X_i, \beta, \alpha) = \text{Prob}(X_i\beta + \varepsilon_i < \alpha_1) = \text{Prob}(\varepsilon_i < \alpha_1 - X_i\beta) = P_{i1}$$
$$\text{Prob}(Y_i = 2 \mid X_i, \beta, \alpha) = \text{Prob}(\alpha_1 \le X_i\beta + \varepsilon_i < \alpha_2) = \text{Prob}(\alpha_1 - X_i\beta \le \varepsilon_i < \alpha_2 - X_i\beta) = P_{i2}$$
$$\text{Prob}(Y_i = 3 \mid X_i, \beta, \alpha) = \text{Prob}(X_i\beta + \varepsilon_i \ge \alpha_3) = \text{Prob}(\varepsilon_i \ge \alpha_3 - X_i\beta) = P_{i3} = 1 - P_{i1} - P_{i2}$$

La vraisemblance de la i^e observation est :

$$L_i = P_{i1}^{\delta i1} \times P_{i2}^{\delta i2} \times (1 - P_{i1} - P_{i2})^{1 - \delta i1 - \delta i2}$$

à condition de poser :

δ_{i1} = 1 si Y_i = 1 et 0 sinon
δ_{i2} = 1 si Y_i = 2 et 0 sinon,

Les coefficients du modèle sont estimés par la méthode du maximum de vraisemblance :

$$\underset{\alpha,\beta}{\text{Max}} \sum_{i=1}^{n} \text{Log } L_i$$

➔ 2 types d'hypothèses retenues pour la distribution du terme d'erreur :

- hypothèse de distribution logistique (logit ordonné)
- hypothèse de normalité (probit ordonné)

Cas du modèle logit ordonné :

$$\hat{P}_{i1} = \Phi\left(\hat{\alpha}_1 - X_i\hat{\beta}\right) = \frac{\exp\left(\hat{\alpha}_1 - X_i\hat{\beta}\right)}{1+\exp\left(\hat{\alpha}_1 - X_i\hat{\beta}\right)}$$

$$\hat{P}_{i2} = \Phi\left(\hat{\alpha}_2 - X_i\hat{\beta}\right) - \Phi\left(\hat{\alpha}_1 - X_i\hat{\beta}\right) = \frac{\exp\left(\hat{\alpha}_2 - X_i\hat{\beta}\right)}{1+\exp\left(\hat{\alpha}_2 - X_i\hat{\beta}\right)} - \frac{\exp\left(\hat{\alpha}_1 - X_i\hat{\beta}\right)}{1+\exp\left(\hat{\alpha}_1 - X_i\hat{\beta}\right)}$$

$$\hat{P}_{i3} = 1 - \hat{P}_{i1} - \hat{P}_{i2}$$

➔ Ces différentes probabilités sont calculées pour chaque individu.

Cas du modèle probit ordonné :

$$\hat{P}_{i1} = \Phi\left(\hat{\alpha}_1 - X_i\hat{\beta}\right) = \int_{-\infty}^{\hat{\alpha}_1 - X_i\hat{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$$

$$\hat{P}_{i2} = \Phi\left(\hat{\alpha}_2 - X_i\hat{\beta}\right) - \Phi\left(\hat{\alpha}_1 - X_i\hat{\beta}\right) = \int_{-\infty}^{\hat{\alpha}_1 - X_i\hat{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz - \int_{-\infty}^{\hat{\alpha}_2 - X_i\hat{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$$

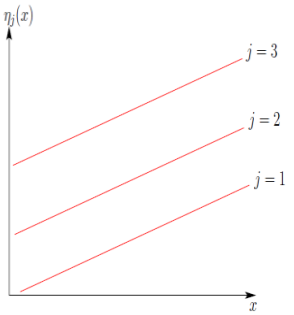
$$= \int_{\hat{\alpha}_1 - X_i\hat{\beta}}^{\hat{\alpha}_2 - X_i\hat{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right) dz$$

$$\hat{P}_{i3} = 1 - \hat{P}_{i1} - \hat{P}_{i2}$$

➔ Estimation de p+k-1 coefficients : p pentes et k-1 constantes.

➔ Hypothèse sous-jacente à un modèle ordonné : hypothèse de **l'égalité des pentes** (« proportional odds ratio » ou « modèle à pentes égales »)

➔ **Coefficients des variables explicatives supposés identiques quel que soit le niveau de Y** tandis que la valeur des constantes change. Quelle que soit la modalité j considérée, une variable explicative a donc la même influence sur P(Y≤j / X=x) et donc sur la probabilité P(Y>j / X=x).



➔ Influence de la variable explicative indépendante de la valeur de la modalité.

➔ Calcul d'odd-ratio dans le cas d'un modèle logit ordonné.

Lorsque la valeur de la variable explicative augmente d'une unité :

$$\frac{\text{Prob}(Y \le j/x) / \text{Prob}(Y > j/x)}{\text{Prob}(Y \le j/x') / \text{Prob}(Y > j/x')} = \frac{e^{\hat{\alpha}_i + x\hat{\beta}_1}}{e^{\hat{\alpha}_i + x'\hat{\beta}_1}} = e^{\hat{\beta}_1}$$

Cas de la commande vglm option reverse=FALSE : la valeur obtenue correspond à l'effet d'une augmentation d'une unité de la variable explicative étudiée sur la probabilité d'être dans la catégorie inférieure par rapport aux catégories supérieures

Cas de la commande polr ou vglm option reverse=TRUE :

$$\frac{\text{Prob}(Y > j/x) / \text{Prob}(Y \le j/x)}{\text{Prob}(Y > j/x') / \text{Prob}(Y \le j/x')} = \frac{e^{\hat{\alpha}_i + x\hat{\beta}_2}}{e^{\hat{\alpha}_i + x'\hat{\beta}_2}} = e^{\hat{\beta}_2}$$

Sachant : $\hat{\beta}_2 = -\hat{\beta}_1$

Odd-ratio : effet d'une augmentation d'une unité de la variable explicative étudiée sur la probabilité d'être dans la catégorie supérieure par rapport aux catégories inférieures.

Effet marginal d'une variable continue : $\dfrac{\partial \text{Prob}(Y=j)}{\partial X_p}$

2) Application sous R

Objectif : analyser les facteurs influençant la possibilité qu'un étudiant de l'université s'inscrive en 2ème cycle

Modalités de Y : peu plausible (1) < probable (2) < très probable (3)).
Variables explicatives :

- Niveau de scolarité des parents (1 si au moins des parents a un diplôme d'université, 0 sinon),
- Etablissement actuel de l'étudiant est un établissement privé ou public (codé 1 si public, 0 sinon),
- Moyenne de ses notes actuelles allant de 1 à 5 (1 la note la plus mauvaise, 5 la meilleure)

str(Master)
'data.frame': 400 obs. of 5 variables:
Nobs : num 1 2 3 4 5 6 7 8 9 10 ...
Parent : num 0 1 1 0 0 0 0 0 0 1 ...
Public : num 0 0 1 0 0 1 0 0 0 0 ...
Note : num 3.26 3.21 3.94 2.81 2.53 2.59 2.56 2.73 3 3.5 ...
Inscription: num 3 2 1 2 2 1 2 2 1 2 ...

cor(Master[,c("Note","Public","Parent")], use="complete.obs", method = c("spearman"))

	Note	*Public*	*Parent*
Note	*1.0000000*	*0.2125454*	*0.1844373*
Public	*0.2125454*	*1.0000000*	*0.0789744*
Parent	*0.1844373*	*0.0789744*	*1.0000000*

Master$Inscription<-ordered(Master$Inscription)
Master$Public<-as.factor(Master$Public)
Master$Parent<-as.factor(Master$Parent)

str(Master)
'data.frame': 400 obs. of 5 variables:
Nobs : num 1 2 3 4 5 6 7 8 9 10 ...
Parent : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 1 1 2 ...
Public : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 1 1 1 ...
Note : num 3.26 3.21 3.94 2.81 2.53 2.59 2.56 2.73 3 3.5 ...
Inscription: Ord.factor w/ 3 levels "1"<"2"<"3": 3 2 1 2 2 1 2 2 1 2 ...

➔ Création d'un tableau croisé entre variables qualitatives

ftable(xtabs(~Inscription+Public+Parent,data=Master))

		Parent	0	1
Inscription	Public			
1	0		175	14
	1		25	6
2	0		98	26
	1		12	4
3	0		20	10
	1		7	3

A) <u>Estimation du modèle ordonné par la fonction polr</u>

library(MASS)
model<-polr(Inscription~Parent+Public+Note,data=Master,method=c("logistic"))
summary(model)

Coefficients:

	Value	Std. Error	t value
Parent[T.1]	1.04769	0.2658	3.9418
Public[T.1]	-0.05879	0.2979	-0.1974
Note	0.61594	0.2606	2.3632

Intercepts:

	Value	Std. Error	t value
1|2	2.2039	0.7795	2.8272
2|3	4.2994	0.8043	5.3453

Residual Deviance: 717.0249
AIC: 727.0249

➔ Rajouter les p-values pour faciliter la lecture

(ctable<-coef(summary(model)))
p <- pnorm(abs(ctable[,"t value"]),lower.tail=FALSE)*2
p2<-round(p,4)
(ctable<-cbind(ctable,pvalue=p2))

	Value	*Std. Error*	*t value*	*pvalue*
Parent[T.1]	*1.04769011*	*0.2657894*	*3.9418053*	*0.0001*
Public[T.1]	*-0.05878572*	*0.2978611*	*-0.1973595*	*0.8435*
Note	*0.61594057*	*0.2606337*	*2.3632420*	*0.0181*
1|2	*2.20391472*	*0.7795447*	*2.8271820*	*0.0047*
2|3	*4.29936313*	*0.8043260*	*5.3452991*	*0.0000*

➔ Notes obtenues par l'étudiant + Avoir au moins un des 2 parents ayant fait des études supérieures : impact de manière significative sur le fait que l'étudiant s'inscrive en master.

➔ Coefficient d'une variable explicative positif : l'inscription peu probable de l'étudiant en master est d'autant plus faible que la valeur de la variable est élevée. Inversement, celle d'une inscription très probable est d'autant plus forte que la valeur de cette variable est élevée.

➔ Un coefficient d'une variable explicative positif signifie que tout accroissement de la variable contribue à rendre plus probable les modalités les plus élevées de la variable à expliquer (ici : l'inscription en master).

➔ Inversement, si le coefficient est négatif, tout accroissement de la variable contribue à rendre moins probable les modalités les plus élevées.

➔ Le fait d'avoir un parent diplômé et d'avoir de meilleures notes rendent plus probable l'inscription en master.

➔ Les valeurs seuils n'ont d'intérêt que pour leur significativité et indiquent que le découpage entre les catégories 1/2/3 ont du sens

➔ Compte tenu du fait que la variable Public n'est pas significative, on peut utiliser la procédure step

step(model, direction='forward', criterion='AIC')
step(model, direction='backward', criterion='AIC')
step(model, direction='both', criterion='AIC')

➔ 2 des 3 procédures aboutissent au même résultat : on ne conserve donc pour le modèle à estimer la variable Note et Parent

➔ On ré-estime donc le modèle

```
model2<-polr(Inscription~Parent+Note,data=Master,method=c("logistic"))
summary(model2)
(ctable<-coef(summary(model2)))
p <- pnorm(abs(ctable[,"t value"]),lower.tail=FALSE)*2
p2<-round(p,4)
(ctable<-cbind(ctable,pvalue=p2))
```


	Value	*Std. Error*	*t value*	*pvalue*
Parent[T.1]	*1.0457078*	*0.2656427*	*3.936520*	*0.0001*
Note	*0.6042468*	*0.2539454*	*2.379436*	*0.0173*
1|2	*2.1762687*	*0.7670896*	*2.837046*	*0.0046*
2|3	*4.2715846*	*0.7921502*	*5.392392*	*0.0000*

➔ Calcul des odds ratios

exp(coef(model2))

Parent[T.1]	*Note*
2.845412	*1.829873*

➔ Variable binaire (passage de 0 à 1)

Les étudiants ayant des parents diplômés ont 2,8 fois plus de chance d'avoir une inscription très probable par rapport aux 2 autres situations (probable, peu probable). De même, les étudiants ayant des parents diplômés ont 2,8 fois plus de chance d'avoir une inscription probable ou très probable par rapport au fait de ne pas être inscrit en master.

➔ Variable quantitative : augmentation d'une unité

Les étudiants dont la note augmente d'un point ont 1,8 fois plus de chance d'avoir une inscription très probable par rapport aux 2 autres situations. Idem Très probable+Probable / peu probable.

confint(model2)

	Estimate	*2.5 %*	*97.5 %*
Parent[T.1]	*1.0457078*	*0.5250576*	*1.566358*
Note	*0.6042468*	*0.1065230*	*1.101971*

➔ Intervalle de confiance sur les odd-ratios

exp(cbind(Estimate=coef(model2),confint(model2)))

	Estimate	*2.5 %*	*97.5 %*
Parent[T.1]	*2.845412*	*1.690556*	*4.789174*
Note	*1.829873*	*1.112404*	*3.010092*

➔ Calcul des résultats de prédictions en termes de probabilité

(m2.pred<-predict(model2, type = "p"))

	1	*2*	*3*
1	*0.5514235*	*0.3575975*	*0.09097896*
2	*0.3080884*	*0.4754225*	*0.21648915*
3	*0.2226704*	*0.4768771*	*0.30045245*
4	*0.6173547*	*0.3117918*	*0.07085357*
5	*0.6564523*	*0.2830554*	*0.06049229*
6	*0.6482303*	*0.2891838*	*0.06258589*

...

➔ En termes de modalité (1 ou 2 ou 3) de la variable inscription

Master$predict.model2 <- predict(model2)

(mc<-table(Master$predict.model2 , Master$Inscription))

	1	*2*	*3*
1	*201*	*110*	*27*
2	*19*	*30*	*13*
3	*0*	*0*	*0*

➔ Qualité en termes de prévision

qualite<-((mc[1,1]+mc[2,2]+mc[3,3])/sum(mc))*100
print(qualite)

[1] 57.75

➔ Qualité du modèle en termes d'ajustement du modèle

Model0<-polr(Inscription~1,data=Master,method=c("logistic"))
R2_Mc_Fadden<-1-(model2$deviance/Model0$deviance)
R2_Mc_Fadden

[1] 0.03257059

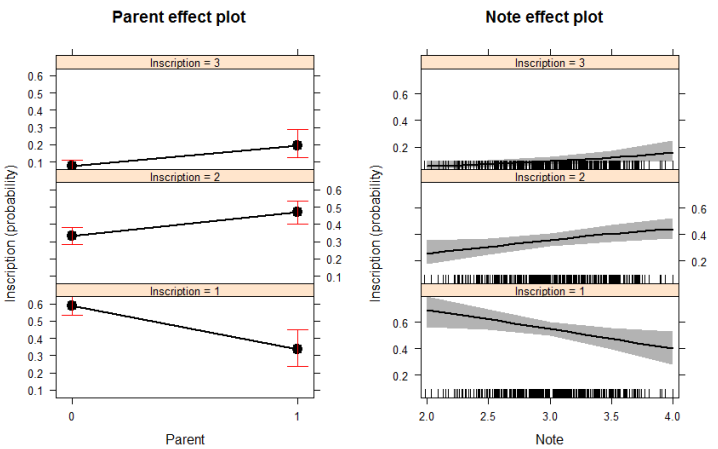
library(lmtest)
lrtest(model2,Model0)

Likelihood ratio test

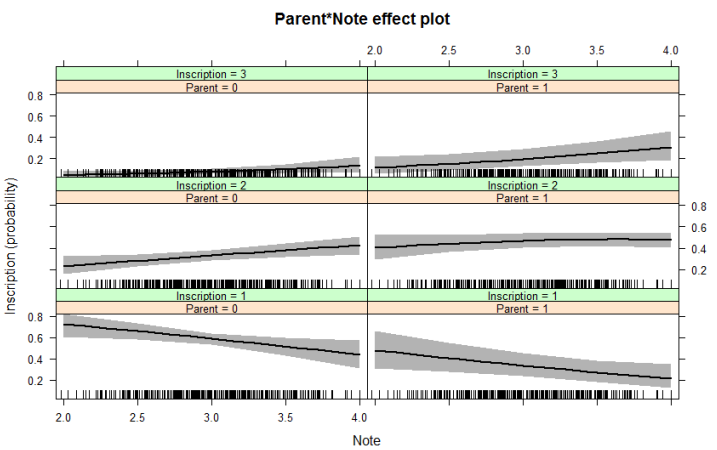
Model 1: Inscription ~ Parent + Note
Model 2: Inscription ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 4 -358.53
2 2 -370.60 -2 24.142 5.725e-06 ***
---
*Signif. codes: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1*

➔ Ce test est équivalent au ratio de vraisemblance dans le cas du modèle (logit) binaire
Comme la p-value < 0,05, cela a un intérêt d'estimer le model2 au seuil de risque de 1%

library(effects)
plot(allEffects(model2))

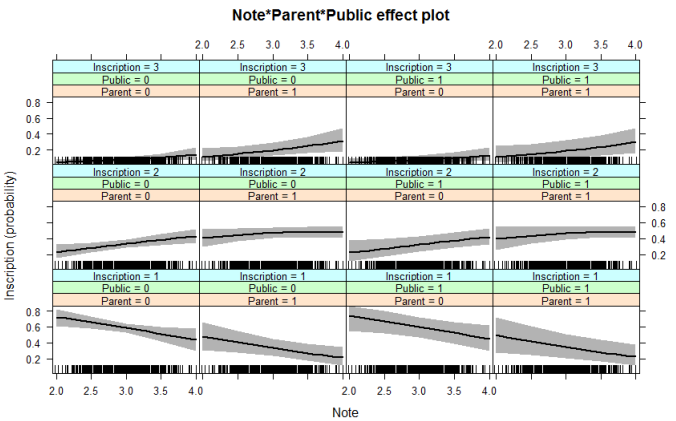


plot(Effect(c("Parent","Note"),model2))



➔ Absence d'effet de la variable Public visible sur le graphique

model1<-polr(Inscription~Parent+Note+Public,data=Master,method=c("logistic"))
plot(Effect(c("Note","Parent","Public"),model1))



<u>Vérification de l'hypothèse de l'égalité des pentes pour la variable Parent</u>

glm(I(as.numeric(Inscription)>=2)~Parent,family="binomial",data=Master)

(Intercept)	*Parent[T.1]*
-0.3783	*1.1438*

Degrees of Freedom: 399 Total (i.e. Null); 398 Residual
Null Deviance : 550.5
Residual Deviance: 534.1 AIC: 538.1

glm(I(as.numeric(Inscription)>=3)~Parent,family="binomial",data=Master)

(Intercept)	*Parent[T.1]*
-2.441	*1.094*

Degrees of Freedom: 399 Total (i.e. Null); 398 Residual
Null Deviance: 260.1
Residual Deviance: 252.2 AIC: 256.2

Vérification de l'hypothèse de l'égalité des pentes pour la variable Note

glm(I(as.numeric(Inscription)>=2)~Note,family="binomial",data=Master)

(Intercept)	*Note*
-2.2084	*0.6683*

Degrees of Freedom: 399 Total (i.e. Null); 398 Residual
Null Deviance: 550.5
Residual Deviance: 543.7 AIC: 547.7

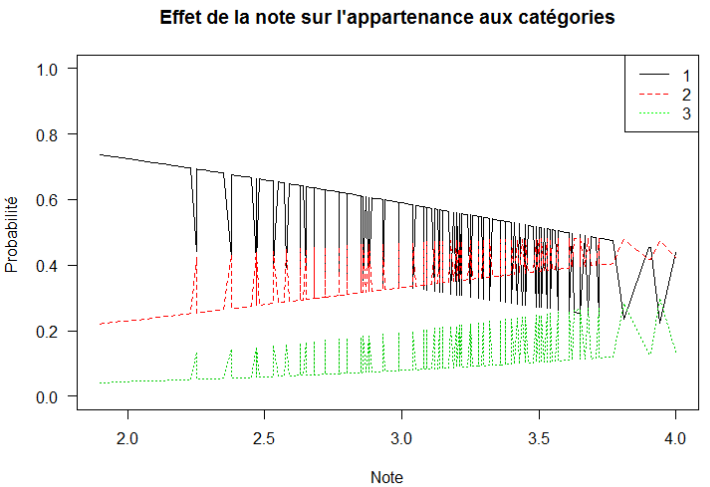
glm(I(as.numeric(Inscription)>=3)~Note,family="binomial",data=Master)

(Intercept)	*Note*
-5.481	*1.071*

Degrees of Freedom: 399 Total (i.e. Null); 398 Residual
Null Deviance : 260.1
Residual Deviance: 253.8 AIC: 257.8

➔ Graphique de l'effet de la variable Note sur la probabilité d'appartenir aux 3 categories définies pour la variable Inscription

```
ooo <- with(Master, order(Note))
m2.pred<-predict(model2, Master, type = "p")
(m2.pred[ooo,])
with(Master, matplot(Note[ooo], m2.pred[ooo,], ylim = c(0,1),
                     xlab = "Note", ylab = "Probabilité ", las = 1,
                     main = "Effet de la note sur l'appartenance aux catégories", type = "l",
                     lwd = 1))
legend("topright",col = c("black","red","green"), lty=1:3, legend=colnames(m2.pred))
```



➔ Toutes les fonctions vues précédemment existent également pour le modèle probit ordonné hormis pour les odd-ratios (fonction avec exp)

```
modelprobit<-polr(Inscription~Parent +Note,data=Master,method=c("probit"))
p <- pnorm(abs(ctable[,"t value"]),lower.tail=FALSE)*2
p2<-round(p,4)
(ctable<-cbind(ctable,pvalue=p2))
```

	Value	*Std.Error*	*t value*	*pvalue*
Parent1	*0.5984099*	*0.1577813*	*3.792653*	*0.0001*
Note	*0.3602627*	*0.1526189*	*2.360537*	*0.0182*
1|2	*1.3017945*	*0.4598727*	*2.830771*	*0.0046*
2|3	*2.5077288*	*0.4693103*	*5.343434*	*0.0000*

B) <u>Estimation du modèle via la fonction vglm</u>

```
library(VGAM)
fit1 <- vglm(Inscription ~ Parent + Public + Note, data=Master, link="logit", family =
cumulative(parallel=TRUE, reverse=TRUE))
summary(fit1)
```

Pearson residuals:

	Min	1Q	Median	3Q	Max
logit(P[Y>=2])	-1.8059	-0.8247	-0.6756	1.1309	1.671
logit(P[Y>=3])	-0.7532	-0.4554	-0.2040	-0.1806	4.057

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept):1	-2.20335	0.78440	-2.809	0.00497 **
(Intercept):2	-4.29879	0.80915	-5.313	1.08e-07 ***
Parent1	1.04766	0.26845	3.903	9.52e-05 ***
Public1	-0.05867	0.28861	-0.203	0.83891
Note	0.61575	0.26258	2.345	0.01903 *

Number of linear predictors: 2
Names of linear predictors: logit(P[Y>=2]), logit(P[Y>=3])
Residual deviance: 717.0249 on 795 degrees of freedom
Log-likelihood: -358.5124 on 795 degrees of freedom
Number of iterations: 4

Exponentiated coefficients:

Parent1	Public1	Note
2.8509582	0.9430165	1.8510513

confint(fit1)

	2.5 %	97.5 %
(Intercept):1	-3.7407492	-0.6659446
(Intercept):2	-5.8846971	-2.7128844
Parent1	0.5215060	1.5738043
Public1	-0.6243370	0.5069940
Note	0.1011095	1.1303980

exp(cbind(Estimate=coef(fit1),confint(fit1)))

	Estimate	*2.5 %*	*97.5 %*
(Intercept):1	*0.11043293*	*0.023736314*	*0.51378797*
(Intercept):2	*0.01358498*	*0.002781689*	*0.06634516*
Parent1	*2.85095816*	*1.684562648*	*4.82496892*
Public1	*0.94301649*	*0.535616404*	*1.66029287*
Note	*1.85105131*	*1.106397837*	*3.09688870*

➔ Estimation du modèle précédent sans la variable Public

fit2 <- vglm(Inscription ~ Parent +Note, data=Master, link="logit", family = cumulative(parallel = TRUE, reverse = TRUE))

…..

Coefficients:

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept):1	*-2.1763*	*0.7726*	*-2.817*	*0.00485 ** *
(Intercept):2	*-4.2716*	*0.7976*	*-5.356*	*8.52e-08 *** *
Parent1	*1.0457*	*0.2682*	*3.899*	*9.67e-05 *** *
Note	*0.6043*	*0.2561*	*2.360*	*0.01829 * *

….

Residual deviance: 717.0638 on 796 degrees of freedom
Log-likelihood: -358.5319 on 796 degrees of freedom

➔ L'option parallel=FALSE permet que la valeur des coefficients ne soient pas identiques d'une classe à l'autre et ceux pour toutes les variables explicatives

fit3 <- vglm(Inscription ~ Parent + Note , data=Master , link="logit", family = cumulative(parallel=FALSE, reverse=TRUE))
summary(fit3)

Pearson residuals:

	Min	1Q	Median	3Q	Max
logit(P[Y>=2])	-1.7939	-0.8209	-0.6901	1.1928	1.596
logit(P[Y>=3])	-0.8145	-0.4325	-0.2088	-0.1738	4.515

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept):1	-1.9969	0.7997	-2.497	0.012519	*
(Intercept):2	-5.2556	1.3816	-3.804	0.000142	***
Parent1:1	1.0635	0.2976	3.574	0.000352	***
Parent1:2	0.9811	0.3770	2.603	0.009246	**
Note:1	0.5431	0.2655	2.045	0.040838	*
Note:2	0.9262	0.4432	2.090	0.036634	*

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

Number of linear predictors: 2
Names of linear predictors: logit(P[Y>=2]), logit(P[Y>=3])

Residual deviance: 716.2281 on 794 degrees of freedom
Log-likelihood: -358.114 on 794 degrees of freedom
Number of iterations: 5

Exponentiated coefficients:

Parent1:1	Parent1:2	Note:1	Note:2
2.896431	2.667518	1.721310	2.524920

➔ Test permettant de vérifier si l'hypothèse d'égalité des pentes est vérifiée

1-pchisq(deviance(fit2)-deviance(fit3),df=df.residual(fit2)-df.residual(fit3))
[1] 0.65845

=> Hypothèse d'égalité des pentes uniquement sur une seule variable

fit4<-vglm(Inscription~ Parent + Note , data=Master , link="logit", family = cumulative(parallel = FALSE~1+Note,reverse=TRUE))
head(coef(fit4, matrix = TRUE))

	logit(P[Y>=2])	*logit(P[Y>=3])*
(Intercept)	*-1.9948467*	*-5.2599895*
Parent[T.1]	*1.0366245*	*1.0366245*
Note	*0.5432266*	*0.9218224*

1-pchisq(deviance(fit2)-deviance(fit4),df=df.residual(fit2)-df.residual(fit4))
[1] 0.373878

fit5<-vglm(Inscription~ Parent+Note,data=Master,link="logit",family=cumulative(parallel=FALSE~1+Parent,reverse=TRUE))
head(coef(fit5, matrix = TRUE))

	logit(P[Y>=2])	*logit(P[Y>=3])*
(Intercept)	*-2.1798595*	*-4.2603249*
Parent[T.1]	*1.0651764*	*1.0058213*
Note	*0.6048393*	*0.6048393*

1-pchisq(deviance(fit2)-deviance(fit5),df=df.residual(fit2)-df.residual(fit5))
[1] 0.8780265

➔ Hypothèse de l'égalité des pentes acceptée au seuil de 5% pour l'ensemble des variables explicatives

➔ Sinon (dans le cas contraire) nécessité de modéliser le choix par un nombre plus restreint de catégories ou de dichotomiser la variable à expliquer (mais perte d'informations)

➔ Reprenons le modèle fit2 (variables explicatives Note et Parent)

On peut calculer les effets marginaux pour tous les individus pour toutes les variables explicatives

margeff(fit2)

, , 1

	1	2	3
(Intercept)	*0.5383210*	*-0.18504885*	*-0.35327218*
Parent1	*-0.2586597*	*0.17217800*	*0.08648167*
Note	*-0.1494668*	*0.09949323*	*0.04997352*

, , 2

	1	2	3
(Intercept)	*0.4639235*	*0.26063372*	*-0.7245573*
Parent1	*-0.2229122*	*0.04553924*	*0.1773729*
Note	*-0.1288100*	*0.02631489*	*0.1024951*

…

➔Calcul des résultats de prédictions en termes de probabilité

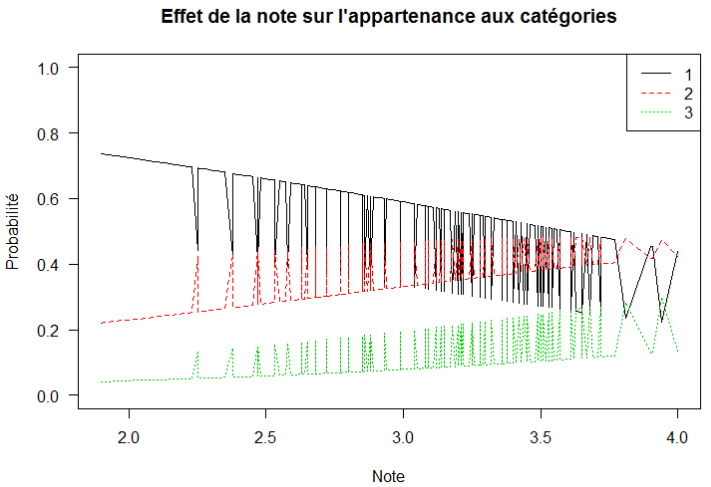
(fitted(fit2))

	1	2	3
1	*0.5514228*	*0.3575977*	*0.09097951*
2	*0.3080897*	*0.4754216*	*0.21648874*
3	*0.2226700*	*0.4768763*	*0.30045370*

...

➔ Graphique de l'effet de la note sur la probabilité d'appartenir à telle ou telle catégorie

```
ooo <- with(Master, order(Note))
fitted(fit2)[ooo,]
with(Master, matplot(Note[ooo], fitted(fit2)[ooo,], ylim = c(0,1),
                     xlab = "Note", ylab = "Probabilité", las = 1,
                     main = "Effet de la note sur l'appartenance aux catégories", type = "l", lwd
                     = 1))
legend("topright",col = c("black","red","green"), lty=1:3, legend=colnames(fitted(fit2)))
```



➔ Qualité du modèle en termes d'ajustement du modèle

fit0 <- vglm(Inscription ~ 1, data=Master, link="logit", family = cumulative(parallel=TRUE, reverse=TRUE))

print(pseudo_R2 <- 1 - deviance(fit2) / deviance(fit0))

[1] 0.03257059

C) <u>Comment calculer les effets marginaux pour chaque modalité de la variable à expliquer et pour chaque variable explicative</u>

Les deux fonctions précédentes polr et vglm sont peu pratiques pour calculer les effets marginaux.

➔ On utilise donc la fonction oglmx

#installer la fonction oglmx
install.packages("oglmx")

library(oglmx)

library(readxl)
Master<-read_excel("Master.xls",sheet="Feuil1",col_names=TRUE)
Master$Public<-as.factor(Master$Public)
Master$Parent<-as.factor(Master$Parent)

➔ Estimation du modèle (avec Parent et Note comme variables explicatives)

results.oprob<-oglmx(Inscription ~ Parent + Note , data=Master,link="logit", constantMEAN=FALSE, constantSD=FALSE, delta=0)
summary(results.oprob)

Ordered Logit Regression

Log-Likelihood: -358.5319
No. Iterations: 6
McFadden's R2: 0.03257059
AIC: 725.0638

	Estimate	Std. error	t value	Pr(>|t|)	
Parent1	*1.04571*	*0.26564*	*3.9365*	*8.267e-05*	***
Note	*0.60425*	*0.25395*	*2.3794*	*0.01734*	*

----- Threshold Parameters -----

	Estimate	Std. error	t value	Pr(>|t|)	
Threshold (1->2)	*2.17628*	*0.76709*	*2.8371*	*0.004553*	**
Threshold (2->3)	*4.27159*	*0.79215*	*5.3924*	*6.952e-08*	***

➔ Calcul des effets marginaux au niveau moyen de l'échantillon (Parent : 0,16 et Note =3)

margins.oglmx(results.oprob, atmeans=TRUE)

Marginal Effects on Pr(Outcome==1)

	Marg.Eff	Std.error	t value	Pr(>|t|)	
Parent1	-0.254132	0.060072	-4.2304	2.332e-05	***
Note	-0.149570	0.062845	-2.3800	0.01731	*

Marginal Effects on Pr(Outcome==2)

	Marg.Eff	Std.Error	t value	Pr(>|t|)	
Parent1	0.137241	0.028503	4.8150	1.472e-06	***
Note	0.099312	0.042797	2.3205	0.02031	*

Marginal Effects on Pr(Outcome==3)

	Marg.Eff	Std Error	t value	Pr(>|t|)	
Parent1	0.116891	0.038684	3.0217	0.002514	**
Note	0.050257	0.021723	2.3135	0.020695	*

➔ Il peut être plus intéressant de calculer les effets marginaux pour Parent = 0 et pour Note = 3)

margins.oglmx(results.oprob, atmeans=TRUE, dummyzero = TRUE)

Marginal Effects on Pr(Outcome==1)

	Marg.Eff	Std.error	t value	Pr(>|t|)	
Parent1	-0.254132	0.060072	-4.2304	2.332e-05	***
Note	-0.146162	0.061563	-2.3742	0.01759	*

Marginal Effects on Pr(Outcome==2)

	Marg.Eff	Std.error	t value	Pr(>|t|)	
Parent1	0.137241	0.028503	4.8150	1.472e-06	***
Note	0.102326	0.043832	2.3345	0.01957	*

Marginal Effects on Pr(Outcome==3)

	Marg.Eff	Std.error	t value	Pr(>|t|)	
Parent1	0.116891	0.038684	3.0217	0.002514	**
Note	0.043836	0.019309	2.2703	0.023191	*

3) Extension du modèle (logit ou probit) ordonné

➔ Les calculs réalisés au préalable supposaient que les erreurs soient homoscédastiques.

La prise en compte de l'hétéroscédasticité des erreurs peut être faite grâce à la fonction oglmx présentée précédemment

➔ On réestime le modèle complet avec les 3 variables explicatives Parent, Note et Public et on suppose que l'hétéroscédasticité des erreurs peut être liée aux 3 variables explicatives

results.oprobhet<-oglmx(Inscription ~ Parent+Note+Public, ~ Parent+Note+Public, data=Master, link="logit", constantMEAN=FALSE, constantSD=FALSE)

summary(results.oprobhet)

Heteroskedastic Ordered Logit Regression
Log-Likelihood: -355.8169
No. Iterations: 21
McFadden's R2: 0.03989639
AIC: 727.6339
----- Mean Equation ------

	Estimate	Std.error	t value	Pr(>|t|)
Parent1	1.82985	1.24022	1.4754	0.1401
Note	0.88613	0.62943	1.4078	0.1592
Public1	-0.61008	0.86511	-0.7052	0.4807

----- SD Equation ------

	Estimate	Std.error	t value	Pr(>|t|)
Parent1	-0.10722	0.19256	-0.5568	0.57764
Note	0.15372	0.22188	0.6928	0.48844
Public1	0.45902	0.24452	1.8772	0.06049 .

----- Threshold Parameters -----

	Estimate	Std.error	t value	Pr(>|t|)
Threshold (1->2)	3.1993	2.1298	1.5022	0.1331
Threshold (2->3)	6.7153	4.2480	1.5808	0.1139

➔ On réestime donc le modèle en supposant que l'hétéroscédasticité des erreurs dépend uniquement de la variable explicative Public

results.oprobhet2<-oglmx(Inscription~Parent+Note+Public,~ Public, data=Master, link="logit", constantMEAN=FALSE, constantSD=FALSE)

summary(results.oprobhet2)

Heteroskedastic Ordered Logit Regression

Log-Likelihood: -356.1302
No. Iterations: 9
McFadden's R2: 0.03905099
AIC: 724.2605
----- Mean Equation ------

	Estimate	*Std.error*	*t value*	*Pr(>|t|)*
Parent1	*1.18060*	*0.28928*	*4.0812*	*4.48e-05 ****
Note	*0.61408*	*0.27383*	*2.2426*	*0.02492 **
Public1	*-0.36143*	*0.48204*	*-0.7498*	*0.45338*

----- SD Equation ------

	Estimate	*Std.error*	*t value*	*Pr(>|t|)*
Public1	*0.48865*	*0.24042*	*2.0325*	*0.04211 **

----- Threshold Parameters -----

	Estimate	*Std.error*	*t value*	*Pr(>|t|)*
Threshold (1->2)	*2.19323*	*0.81832*	*2.6802*	*0.007359 ***
Threshold (2->3)	*4.45794*	*0.84762*	*5.2593*	*1.446e-07 ****

➔ La variable explicative Public n'étant pas significative au seuil de 10%, on réestime le modèle sans cette variable

results.oprobhet3<-oglmx(Inscription~Parent+Note,~ Public, data=Master, link="logit", constantMEAN=FALSE, constantSD=FALSE)

summary(results.oprobhet3)

Heteroskedastic Ordered Logit Regression

Log-Likelihood: -356.4468
No. Iterations: 8
McFadden's R2: 0.03819677
AIC: 722.8936
----- Mean Equation ------

	Estimate	Std.error	t value	Pr(>|t|)
Parent1	1.16274	0.28735	4.0464	5.201e-05 ***
Note	0.57965	0.26890	2.1557	0.03111 *

----- SD Equation ------

	Estimate	Std.error	t value	Pr(>|t|)
Public1	0.43582	0.22744	1.9163	0.05533 .

----- Threshold Parameters -----

	Estimate	Std.error	t value	Pr(>|t|)
Threshold (1->2)	2.11083	0.80750	2.6140	0.008948 **
Threshold (2->3)	4.36358	0.83602	5.2195	1.794e-07 ***

➔ Test permettant de comparer le modèle avec et sans hétéroscédasticité des erreurs

library(lmtest)
lrtest(results.oprob,results.oprobhet3)

Likelihood ratio test

Model 1: Inscription ~ Parent + Note
Model 2: Inscription ~ Parent + Note | Public

	Df	LogLik	Df	Chisq	Pr(>Chisq)
1	4	-358.53			
2	5	-356.45	1	4.1702	0.04114 *

➔ Il existe une différence significative au seuil de risque de 5% entre les deux modèles

➔ Il faut donc conserver les résultats du modèle avec la prise en compte de l'hétéroscédascticité des erreurs (notamment pour calculer les effets marginaux)

margins.oglmx(results.oprobhet3,atmeans = TRUE)

Marginal Effects on Pr(Outcome==1)

	Marg.Eff	Std.Error	t value	Pr(>|t|)
Parent1	-0.264210	0.059173	-4.4650	8.006e-06 ***
Note	-0.135114	0.063027	-2.1437	0.03205 *
Public1	-0.016622	0.011309	-1.4698	0.14162

Marginal Effects on Pr(Outcome==2)

	Marg.Eff	Std.error	t value	Pr(>|t|)
Parent1	0.140500	0.027786	5.0564	4.272e-07 ***
Note	0.089816	0.042749	2.1010	0.03564 *
Public1	-0.074242	0.044700	-1.6609	0.09674 .

Marginal Effects on Pr(Outcome==3)

	Marg.Eff	Std.error	t value	Pr(>|t|)
Parent1	0.123710	0.039717	3.1148	0.001841 **
Note	0.045298	0.021692	2.0882	0.036781 *
Public1	0.090864	0.049852	1.8227	0.068352 .

margins.oglmx(results.oprobhet3, atmeans = TRUE, dummyzero = TRUE)

Marginal Effects on Pr(Outcome==1)

	Marg.Eff	Std.error	t value	Pr(>|t|)
Parent1	-0.279945	0.062897	-4.4509	8.552e-06 ***
Note	-0.140000	0.065199	-2.1473	0.03177 *
Public1	-0.032126	0.016611	-1.9340	0.05312 .

Marginal Effects on Pr(Outcome==2)

	Marg.Eff	Std.error	t value	Pr(>|t|)
Parent1	0.159393	0.032364	4.9250	8.438e-07 ***
Note	0.103499	0.048668	2.1266	0.03345 *
Public1	-0.055092	0.037314	-1.4765	0.13982

Marginal Effects on Pr(Outcome==3)

	Marg.Eff	Std.error	t value	Pr(>|t|)
Parent1	0.120552	0.039978	3.0154	0.002566 **
Note	0.036501	0.018105	2.0161	0.043791 *
Public1	0.087218	0.049143	1.7748	0.075934 .

4) Modèle relevant partiellement l'hypothèse d'égalité des pentes (sous Stata) du modèle logit ordonné

Comment importer une base Excel sous Stata 14

import excel "C:\Users\travers-m\Desktop\Cours_2020_2021\Econometrie_variables_qualitatives_M1_EKAP_M2_CODEME\ Cours\Exercice\Master.xls", sheet("Feuil1") firstrow clear

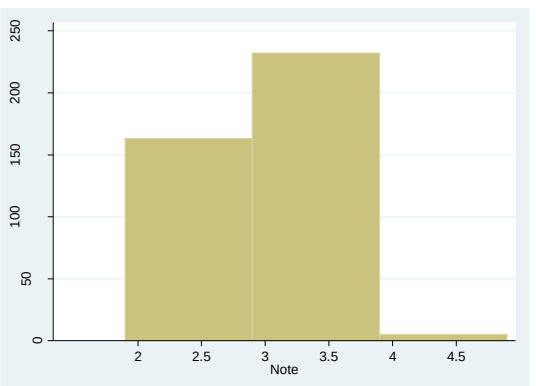
Comment réaliser quelques statistiques :

tab Inscription
tab Inscription Parent
tab Inscription Public
summarize Note
table Inscription, cont(mean Note sd Note)

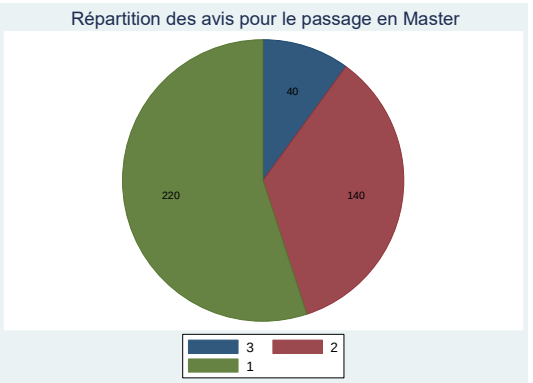
Comment réaliser une box plot

graph box Note

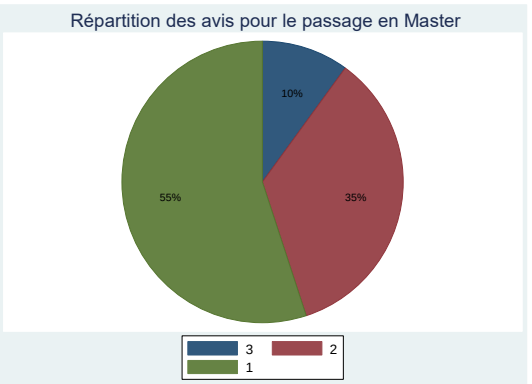
Comment réaliser un histogramme
histogram Note, width(1) frequency



graph pie, over(Inscription) sort plabel(_all sum) title("Répartition des avis pour le passage en Master") legend(on)



graph pie, over(Inscription) sort plabel(_all percent) title("Répartition des avis pour le passage en Master") legend(on)



#Estimation du modèle ordonné avec les 3 variables explicatives Parent Public et Note
ologit Inscription i.Parent i.Public Note
#Calcul des odd ratios pour le modèle ci-dessus
ologit Inscription i.Parent i.Public Note, or
#Vérification au niveau global de l'égalité des pentes pour le modèle ci-dessus
omodel logit Inscription Parent Public Note
Prob > chi2 = 0.2553

➔Puisque la variable Public ne sort pas significative, estimation du modèle sans cette variable
ologit Inscription i.Parent Note
#Vérification de l'égalité des pentes pour le modèle sans la variable explicative Public
omodel logit Inscription Parent Note
Prob > chi2 = 0.6685
➔ On retrouve le résultat de l'estimation réalisé sous R

➔ On s'aperçoit qu'en supprimant la variable Public l'égalité des pentes pour les deux variables (au niveau global) Parent et Note s'améliore ➔ Implique vraisemblablement l'hétéroscédasticité des erreurs : Piste à creuser par la suite

Calcul des effets marginaux (pour le modèle sans Public)
ologit Inscription i.Parent Note
margins, dydx(*) predict(outcome(1)) atmeans at(Parent = 0)
margins, dydx(*) predict(outcome(2)) atmeans at(Parent = 0)
margins, dydx(*) predict(outcome(3)) atmeans at(Parent = 0)

➔ On retrouve les résultats obtenus par la fonction oglmx sous R en supposant l'hypothèse d'homoscédasticité des erreurs avec le modèle Parent et Note

Prise en compte de l'hétéroscédasticité des erreurs (cf démarche oglmx sous R)
oglm Inscription Parent Public Note, hetero(Parent Note Public)
oglm Inscription Parent Note Public, hetero(Public) store(oglm)
oglm Inscription Parent Note, hetero(Public) store(oglm)

Calcul des effets marginaux dans le modèle dont les erreurs sont hétéroscédastiques
margins, dydx(*) predict(outcome(1)) atmeans at(Parent = 0 Public=0)
margins, dydx(*) predict(outcome(2)) atmeans at(Parent = 0 Public=0)
margins, dydx(*) predict(outcome(3)) atmeans at(Parent = 0 Public=0)
➔ On retrouve les résultats obtenus par la fonction oglmx sous R (dans les mêmes conditions)

➔ Il existe sous Stata 14, une fonction qui permet de vérifier de manière automatique les variables ne vérifiant pas l'égalité des pentes et d'estimer un tel modèle

gologit2 Inscription Parent Public Note, autofit(.05)

Testing parallel lines assumption using the .05 level of significance...

Step 1: Constraints for parallel lines imposed for Parent (P Value = 0.8193)
Step 2: Constraints for parallel lines imposed for Note (P Value = 0.7810)
Step 3: Constraints for parallel lines are not imposed for Public (P Value = 0.03252)

Wald test of parallel lines assumption for the final model:

(1) [1]Parent - [2]Parent = 0
(2) [1]Note - [2]Note = 0

chi2(2) = 0.13
Prob > chi2 = 0.9369

An insignificant test statistic indicates that the final model does not violate the proportional odds/ parallel lines assumption
If you re-estimate this exact same model with gologit2, instead of autofit you can save time by using the parameter

pl(Parent Note)
…

Ré-estimation du modèle en précisant pl(Parent Note)

gologit2 Inscription Parent Public Note, pl(Parent Note) store(gologit2)

Generalized Ordered Logit Estimates	*Number of obs*	*400*
	Wald chi2(4)	*27.70*
	Prob > chi2	*0.0000*
Log likelihood = -356.57077	*Pseudo R2*	*0.0379*

(1) [1]Parent - [2]Parent = 0
(2) [1]Note - [2]Note = 0

Inscription	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
Parent	1.057633	.2665412	3.97	0.000	.5352216	1.580044
Public	-.2350038	.3052548	-0.77	0.441	-.8332922	.3632847
Note	.6105983	.2607849	2.34	0.019	.0994694	1.121727
_cons	-2.165854	.7798055	-2.78	0.005	-3.694245	-.6374635
2						
Parent	1.057633	.2665412	3.97	0.000	.5352216	1.580044
Public	.5732672	.4106292	1.40	0.163	-.2315513	1.378086
Note	.6105983	.2607849	2.34	0.019	.0994694	1.121727
_cons	-4.410604	.8088948	-5.45	0.000	-5.996009	-2.8252

➔ Les seuils correspondent aux variables _Cons dans chaque catégorie de la variable Inscription.

➔ On remarque que la variable Public n'a pas le même coefficient selon le niveau de la variable Inscription puisque l'hypothèse d'égalité des pentes n'est pas vérifiée. Son effet augmente et devient positif pour la catégorie Inscription=2

➔ Néanmoins, cette variable n'a pas d'impact significatif au seuil de 10%

Sous Stata, il est possible de vérifier si la forme fonctionnelle utilisée pour le modèle précédent est correct

linktest

Generalized Ordered Logit Estimates — Number of obs = 400
LR chi2(4) = 27.95
Prob > chi2 = 0.0000
Log likelihood = -356.62616 — Pseudo R2 = 0.0377

Inscription	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
_hat	.9887635	.2315844	4.27	0.000	.5348664	1.442661
_hatsq	-.1681605	.4026591	-0.42	0.676	-.9573579	.6210369
_cons	.0409296	.1631399	0.25	0.802	-.2788188	.3606781
2						
_hat	.6016029	.3400144	1.77	0.077	-.0648131	1.268019
_hatsq	1.097684	.6194379	1.77	0.076	-.1163918	2.31176
_cons	-2.48296	.2907551	-8.54	0.000	-3.052829	-1.91309

Pour vérifier si cela un intérêt de faire ce modèle par rapport à celui qui suppose l'égalité des pentes pour l'ensemble des variables

gologit2 Inscription Parent Public Note,pl store(gologit)
lrtest gologit gologit2

Likelihood-ratio test *LR chi2(1) = 3.88*
(Assumption: gologit nested in gologit2) *Prob > chi2 = 0.0488*

➔ Au seuil de risque de 5%, on ne peut conserver le modèle où toutes les variables explicatives sont supposées suivre l'égalité des pentes.

➔Réestimer le modèle en ne prenant pas en compte la variable Public car p-value = 0,84➔ On retombe ainsi sur les résultats de la procédure faite sous R

➔ Au seuil de risque de 1%, on peut prendre le modèle de la slide 61 où le modèle ne prend en compte que partiellement l'égalité des pentes pour les variables Parent et Note

➔ Le calcul des odd ratios se fait de la manière suivante :

gologit2 Inscription Parent Public Note, or pl(Parent Note) store(gologit2)

```
Generalized Ordered Logit Estimates             Number of obs     =        400
                                                Wald chi2(4)      =      27.70
                                                Prob > chi2       =     0.0000
Log likelihood = -356.57077                     Pseudo R2         =     0.0379

 ( 1)  [1]Parent - [2]Parent = 0
 ( 2)  [1]Note - [2]Note = 0
```

Inscription	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1						
Parent	2.879546	.7675177	3.97	0.000	1.707827	4.855169
Public	.7905679	.2413246	-0.77	0.441	.4346161	1.438045
Note	1.841533	.480244	2.34	0.019	1.104585	3.070153
_cons	.114652	.0894062	-2.78	0.005	.0248662	.5286316
2						
Parent	2.879546	.7675177	3.97	0.000	1.707827	4.855169
Public	1.774054	.7284782	1.40	0.163	.793302	3.967299
Note	1.841533	.480244	2.34	0.019	1.104585	3.070153
_cons	.0121478	.0098263	-5.45	0.000	.0024887	.0592968

➔ Pour ce type de modèle, on peut également calculer les effets marginaux

margins, dydx(*) predict(outcome(1)) atmeans at(Parent = 0 Public=0)

```
Conditional marginal effects                    Number of obs     =        400
Model VCE    : OIM

Expression   : Pr(Inscription==1), predict(outcome(1))
dy/dx w.r.t. : Parent Public Note
at           : Parent          =           0
               Public          =           0
               Note            =    2.998925 (mean)
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parent	-.2571386	.0633093	-4.06	0.000	-.3812226	-.1330546
Public	.0571357	.0746001	0.77	0.444	-.0890778	.2033491
Note	-.1484527	.0637441	-2.33	0.020	-.2733889	-.0235164

margins, dydx(*) predict(outcome(2)) atmeans at(Parent = 0 Public=0)

```
Conditional marginal effects                    Number of obs     =        400
Model VCE    : OIM

Expression   : Pr(Inscription==2), predict(outcome(2))
dy/dx w.r.t. : Parent Public Note
at           : Parent          =           0
               Public          =           0
               Note            =    2.998925 (mean)
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parent	.1878579	.0510502	3.68	0.000	.0878013	.2879144
Public	-.0946878	.0660318	-1.43	0.152	-.2241078	.0347322
Note	.1084551	.0470538	2.30	0.021	.0162313	.200679

margins, dydx(*) predict(outcome(3)) atmeans at(Parent = 0 Public=0)

```
Conditional marginal effects                    Number of obs     =        400
Model VCE    : OIM

Expression   : Pr(Inscription==3), predict(outcome(3))
dy/dx w.r.t. : Parent Public Note
at           : Parent          =           0
               Public          =           0
               Note            =    2.998925 (mean)
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Parent	.0692807	.0175751	3.94	0.000	.0348341	.1037274
Public	.0375521	.0247446	1.52	0.129	-.0109464	.0860506
Note	.0399975	.0184827	2.16	0.030	.0037721	.076223

➔ Cette fonction (ologit2) ne permet pas de prendre en compte le cas de l'hétéroscédascité des erreurs mais comme la prise en compte de la non-vérification de l'égalité des pentes (ici Public) implique un effet différent sur les modalités de Y, elle peut avoir un impact significatif sur la variance du modèle

➔En pratique, les résultats des deux estimations suivantes sont comparés

gologit2 Inscription Parent Public Note, pl(Parent Note)
Pseudo R2 = 0.0379

oglm Inscription Parent Note, hetero(Public)
Pseudo R2 = 0.0382

➔ *Le Pseudo R2 sont quasi identiques, le choix peut se faire en fonction de la facilité d'interprétation (le 2ème est plus facile à faire comprendre à des décideurs publics)*

```
Heteroskedastic Ordered Logistic Regression     Number of obs      =        400
                                                LR chi2(3)         =      28.31
                                                Prob > chi2        =     0.0000
Log likelihood = -356.44682                     Pseudo R2          =     0.0382
```

Inscription	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Inscription						
Parent	1.162737	.2873511	4.05	0.000	.5995394	1.725935
Note	.57965	.268897	2.16	0.031	.0526215	1.106679
lnsigma						
Public	.435825	.2274362	1.92	0.055	-.0099418	.8815918
/cut1	2.110829	.8075036	2.61	0.009	.528151	3.693507
/cut2	4.363584	.8360186	5.22	0.000	2.725017	6.00215

Remarque : on peut également utiliser la fonction ologit2 pour estimer un modèle logit

recode Inscription (1 = 0)(2 3 = 1), gen(Passage)
gologit2 Passage Note Public Parent

```
Generalized Ordered Logit Estimates             Number of obs      =        400
                                                LR chi2(3)         =      20.59
                                                Prob > chi2        =     0.0001
Log likelihood =  -264.9624                     Pseudo R2          =     0.0374
```

Passage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Note	.5482457	.2724341	2.01	0.044	.0142846	1.082207
Public	-.2005571	.3053354	-0.66	0.511	-.7990035	.3978894
Parent	1.059612	.2973854	3.56	0.000	.4767471	1.642476
_cons	-1.982971	.812215	-2.44	0.015	-3.574883	-.3910588

gologit2 Passage Note Public Parent, or

```
Generalized Ordered Logit Estimates             Number of obs      =        400
                                                LR chi2(3)         =      20.59
                                                Prob > chi2        =     0.0001
Log likelihood =  -264.9624                     Pseudo R2          =     0.0374
```

Passage	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Note	1.730215	.4713696	2.01	0.044	1.014387	2.951185
Public	.8182748	.2498483	-0.66	0.511	.4497769	1.488679
Parent	2.885251	.8580314	3.56	0.000	1.610826	5.167952
_cons	.1376597	.1118092	-2.44	0.015	.0280187	.6763404

linktest

```
Generalized Ordered Logit Estimates             Number of obs      =        400
                                                LR chi2(2)         =      21.04
                                                Prob > chi2        =     0.0000
Log likelihood = -264.73681                     Pseudo R2          =     0.0382
```

Passage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_hat	1.064286	.2486275	4.28	0.000	.5769853	1.551587
_hatsq	-.3032046	.4518648	-0.67	0.502	-1.188843	.5824342
_cons	.0881169	.1738162	0.51	0.612	-.2525566	.4287905

margins, dydx(*) predict(outcome(1)) atmeans at(Parent = 0 Public=0)

```
Conditional marginal effects                    Number of obs     =        400
Model VCE    : OIM

Expression   : Pr(Passage==1), predict(outcome(1))
dy/dx w.r.t. : Note Public Parent
at           : Note            =    2.998925 (mean)
               Public          =           0
               Parent          =           0
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Note	.1332018	.0665423	2.00	0.045	.0027812	.2636224
Public	-.0487274	.0745203	-0.65	0.513	-.1947844	.0973297
Parent	.2574433	.0706692	3.64	0.000	.1189342	.3959524

5) Complément : Hypothèse d'égalité des pentes partiellement relevée sous R : fonction clm (library(ordinal))

On peut également relever l'hypothèse d'égalité des pentes avec le logiciel R

library(readxl)
Master<-read_excel("Master.xls",sheet="Feuil1",col_names=TRUE)
Master$Inscription<-ordered(Master$Inscription)
Master$Parent<-as.factor(Master$Parent)
Master$Public<-as.factor(Master$Public)

➔ Utilisation de la fonction clm du package ordinal

#Estimation du modèle ordonné avec hypothèse d'égalité des pentes (avec homoscédasticité des erreurs)

fm1<-clm(Inscription~Parent+Note+Public,data=Master)
summary(fm1)

link	*threshold*	*nobs*	*logLik*	*AIC*	*niter*	*max.grad*	*cond.H*
logit	*flexible*	*400*	*-358.51*	*727.02*	*5(0)*	*1.63e-10*	*1.3e+03*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
Parent1	*1.04766*	*0.26579*	*3.942*	*8.09e-05*	***
Note	*0.61575*	*0.26063*	*2.363*	*0.0182*	*
Public1	*-0.05868*	*0.29786*	*-0.197*	*0.8438*	

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

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	*2.2033*	*0.7795*	*2.826*
2|3	*4.2988*	*0.8043*	*5.345*

exp(fm1$beta)

Parent1	Note	Public1
2.8509825	1.8510366	0.9430059

➔ On vérifie l'hypothèse d'égalité des pentes pour les 3 variables explicatives avec le test ci-dessous

nominal_test(fm1)

Tests of nominal effects

formula: Inscription ~ Parent + Note + Public

	Df	logLik	AIC	LRT	Pr(>Chi)	
<none>		*-358.51*	*727.02*			
Parent	*1*	*-358.50*	*729.00*	*0.0247*	*0.87510*	
Note	*1*	*-358.11*	*728.22*	*0.8016*	*0.37061*	
Public	*1*	*-356.57*	*725.14*	*3.8833*	*0.04877*	*

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

➔ Au seuil de risque de 10 % (ici 5%), il n'y a que la variable explicative Public qui ne vérifie pas l'hypothèse d'égalité des pentes.

➔ On estime donc le modèle suivant tenant compte de ce constat

fm2<-clm(Inscription~Parent+Note,nominal=~Public,data=Master)
summary(fm2)

link threshold nobs logLik AIC niter max.grad cond.H
Logit flexible 400 -356.57 725.14 5(0) 2.44e-09 1.3e+03

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
Parent1	*1.0576*	*0.2665*	*3.968*	*7.25e-05*	***
Note	*0.6106*	*0.2608*	*2.341*	*0.0192*	*

Threshold coefficients:

	Estimate	Std. Error	z value
1|2.(Intercept)	*2.1659*	*0.7798*	*2.777*
2|3.(Intercept)	*4.4106*	*0.8089*	*5.453*
1|2.Public1	*0.2350*	*0.3053*	*0.770*
2|3.Public1	*-0.5733*	*0.4106*	*-1.396*

➔ Calcul du R^2 Mac Fadden associé au modèle fm2
fm0<-clm(Inscription~1,data=Master)
llnull<-fm0$logLik
ll<-fm2$logLik
(R2<-1-(ll/llnull))
[1] 0.03786231

➔ On peut vérifier également si les erreurs du modèle fm1 sont homoscédastiques ou non

scale_test(fm1)

Tests of scale effects

formula: Inscription ~ Parent + Note + Public

	Df	logLik	AIC	LRT	Pr(>Chi)	
<none>		*-358.51*	*727.02*			
Parent	*1*	*-358.49*	*728.98*	*0.0456*	*0.83090*	
Note	*1*	*-357.95*	*727.90*	*1.1216*	*0.28957*	
Public	*1*	*-356.13*	*724.26*	*4.7644*	*0.02905*	*

➔ Estimation du modèle tenant compte de l'hétéroscédasticité des erreurs (avec hypothèse d'égalité des pentes) associée à la variable Public

fm3<-clm(Inscription~Parent+Note,scale=~Public,data=Master)
summary(fm3)

link threshold nobs logLik AIC niter max.grad cond.H
logit flexible 400 -356.45 722.89 8(0) 1.53e-10 1.3e+03

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
Parent1	*1.1627*	*0.2874*	*4.046*	*5.2e-05*	***
Note	*0.5796*	*0.2689*	*2.156*	*0.0311*	*

log-scale coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
Public1	*0.4358*	*0.2274*	*1.916*	*0.0553*	.

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	*2.1108*	*0.8075*	*2.614*
2|3	*4.3636*	*0.8360*	*5.219*

fm0<-clm(Inscription~1,data=Master)
llnull<-fm0$logLik
ll<-fm3$logLik
(R2<-1-(ll/llnull))
[1] 0.03819677

➔ Rappel : le Pseudo R2 pour le modèle fm2 est 0.0378. Les Pseudo R2 sont quasi identiques, le choix peut se faire en fonction de la facilité d'interprétation.

6) Exercice d'application

➔ Analyse des déterminants de la consommation individuelle de vin en France en 2000

Constat :

Consommateur régulier (Presque tous les jours ou tous les jours) : 23,8 %
Consommateur occasionnel (1 à 2 fois par semaine ou plus rarement) : 42,6 %
Jamais (Non consommateur) : 33,6 %

Remarques :

SOCF : variable indicatrice du niveau de sociabilité du foyer construite à partir de la fréquence des bons repas en famille ou avec des invités au domicile (ici : Foyer faible sociabilité =1 versus les autres = 0)

THD : variable construite à partir du nombre de jours de fréquentation de divers lieux de consommation hors domicile : sédentaire : les personnes consomment peu hors de chez elles, nomade : sont à l'extérieur presque tous les jours, habitué de bar, convive : fréquentent les foyers amis, rationnaire : mangent régulièrement sur le lieu de travail

VALEUR : 1 : avoir de bonnes relations sociales avec les autres (insertion sociale), 2 : profiter de l'existence et se réaliser pleinement (hédonisme), 3 : se sentir en sécurité (sécurité), 4 : avoir le respect de soi (dignité-respectabilité)

RISQUE: différentes perceptions des risques ou des bienfaits liés à la consommation modérée de vin

TFOYBA : niveau de boissons alcoolisés habituellement présente dans le foyer d'enfance

INTERDIT : différentes attitudes de l'entourage à l'égard de la consommation des individus au moment de l'enfance

STUT : variable crée à partir de plusieurs variables (le revenu du ménage, le revenu par unité de consommation dans le ménage, le niveau d'équipement électro-ménager, la CSP du chef de famille, le niveau scolaire du chef de famille.
La projection de chaque ménage sur le premier axe factoriel ➔ Création d'une échelle ordinale de statut sociale divisée en 5 classes

VILLEENF : =1 si la personne habite toujours dans sa ville d'origine, 0 sinon

Analyses :

Via un modèle probit (binaire) (modèle I)
Via un modèle probit ordonné (référence Non consommateur) (modèle II)

Questions :

A partir des documents fournis :

1) Interpréter l'ensemble des résultats du premier tableau
2) Analyser les effets marginaux pour les deux types de modèles
3) Quelle hypothèse est faite dans le cadre du modèle II ?
4) Que doivent également vérifier les auteurs ?

Variables explicatives		Modèle I (Probit simple)		Modèle II (Probit ordonné)	
		Paramètres	Ecart-type	Paramètres	Ecart-type
Constante		-0,369	0,227	-0,973***	0,173
AGE âge		0,077***	0,007	0,065***	0,006
AGECARRE âge au carré		-0,001***	0,000	-0,0004***	0,000
SIT	= 1 Non maritale	-0,190***	0,068	-0,176***	0,055
PLOGT	= 1 Propriétaire logement	0,050	0,056	0,106***	0,041
THD	= 1 Sédentaire/rationnaire	Référence		Référence	
	= 2 Convive	-0,111*	0,061	0,134***	0,045
	= 3 Habitué de bar	0,072	0,089	0,233***	0,065
	= 4 Nomade	0,077	0,222	0,291***	0,135
INTERDIT	= 1 Aucun interdit	Référence		Référence	
	= 2 Interdit hors vin	0,234***	0,066	0,106**	0,045
	= 3 Vin interdit	0,041	0,055	0,018	0,042
REC	= 1 Ouest	Référence		Référence	
	= 2 Nord	-0,354	0,104	-0,281***	0,080
	= 3 Est	-0,301***	0,103	-0,293***	0,073
	= 4 Nord-Ouest	-0,155	0,106	-0,238***	0,074
	= 5 Sud-Est	-0,229***	0,095	-0,213***	0,066
	= 6 Centre	-0,054***	0,109	-0,086	0,075
	= 7 Sud	-0,155***	0,115	-0,095	0,080
	= 8 Ile-de-France	-0,405***	0,092	-0,301***	0,066
STUT	= 1 Très modeste	-0,675***	0,132	-0,539***	0,092
	= 2 Modeste	-0,435***	0,113	-0,270***	0,065
	= 3 Moyen inférieur	-0,416***	0,112	-0,319***	0,064
	= 4 Moyen supérieur	-0,225***	0,107	-0,197***	0,056
	= 5 Aisé	Référence		Référence	
NPF	= 1 Nombre	0,533***	0,103	0,514***	0,081
	= 2 de personnes	0,219***	0,082	0,342***	0,062
	= 3 au foyer	0,221***	0,081	0,161***	0,062
	= 4	0,231***	0,081	0,066	0,060
	= 5	Référence		Référence	
RISQUE	= 1 Vin préventif	Référence		Référence	
	= 2 Préventif et risqué	-0,113	0,084	-0,039	0,056
	= 3 Sans effet	-0,267***	0,064	-0,155***	0,045
	= 4 Pas d'opinion	-0,394***	0,081	-0,329***	0,064
	= 5 Risqué	-0,477***	0,079	-0,497***	0,062
SEXE	= 1 Masculin	0,372***	0,050	0,609***	0,037
SOCF	= 1 Foyer faible « sociabilité »	-0,264***	0,059	-0,400***	0,050
TFOYBA	= 0 Pas de boisson alcoolisée	-0,776***	0,081	-0,775***	0,080
	= 1 Vin quotidien	Référence		Référence	
	= 2 Autre b. a. quotidienne	-0,166*	0,090	-0,160**	0,066
	= 3 Vin hebdomadaire	0,075	0,079	-0,082	0,050
	= 4 Autres b. a. hebdomadaires	-0,077	0,118	-0,305***	0,089
	= 5 b. a. plus rarement	-0,132*	0,074	-0,422***	0,054
VILLEENF	= 1 Ville d'enfance	-0,069	0,055	-0,026	0,041
VALEUR	= 1 Insertion sociale	-0,110	0,064	-0,031	0,045
	= 2 Hédonisme	Référence		Référence	
	= 3 Sécuritaire	-0,326***	0,086	-0,232***	0,069
	= 4 Dignité, respect de soi	-0,159**	0,069	-0,082*	0,049
SEUIL 1 2e seuil du probit ordonné				0,866***	0,024
SEUIL 2 3e seuil du probit ordonné				1,531***	0,031
SEUIL 3 4e seuil du probit ordonné				1,728***	0,032
Nombre d'observations :		4010		4010	
Nombre de paramètres :		40		43	
Pseudo-R^2 :		0,189		0,146	
LogL :		-1763,433		-5041,000	

Noms symboliques	Modalités	Probit binaire (Modèle I)	Probit ordonné simple (Modèle II)		
		Consommer	Réguliers	Occasionnels	Pas de consommation
Probabilités moyennes calculées (fréquences observées)		0,7670(0,766)	0,231 (0,23)	0,43(0,44)	0,32(0,33)
AGE	Age (c	0,112***	0,211***	0,059***	-0,271***
SIT =1	Non maritale	-0,052***	-0,037***	-0,013***	0,050***
PLOGT=1	Propriétaire logement	0,011***	0,029***	-0,003***	-0,026***
THD =2	Convive	-0,028***	0,030***	0,007***	-0,038***
THD =3	Habitué de bar	0,018***	0,053***	0,014***	-0,067***
THD =4	Nomade	0,012***	0,093***	-0,030***	-0,063***
INTERDIT=2	Interdit sauf vin	0,053***	0,027***	0,002**	-0,028***
INTERDIT=3	Vin interdit	0,011***	0,004***	0,001***	-0,005***
REC=2	Nord	-0,091***	-0,066***	-0,014***	0,080***
REC=3	Est	-0,07***	-0,068***	-0,013***	0,081***
REC=4	Nord-Ouest	-0,037***	-0,058***	-0,010***	0,068***
REC=5	Sud-Est	-0,051***	-0,054***	-0,002	0,056***
REC=6	Centre	-0,011***	-0,023***	0,001	0,022***
REC=7	Sud	-0,034***	-0,02***	0,002*	0,024***
REC=8	Ile-de-France	-0,098***	-0,072***	-0,010***	0,081***
STUT=1	Foyer très modeste	-0,182***	-0,127***	-0,021***	0,148***
STUT=2	Foyer modeste	-0,102***	-0,068***	-0,001	0,069***
STUT=3	Foyer moyen inférieur	-0,094***	-0,077***	-0,008***	0,085***
STUT=4	Foyer moyen supérieur	-0,046***	-0,051***	-0,003**	0,053***
NPF =1	Nombre de personnes au foyer	0,152***	0,121***	0,031***	-0,152***
NPF =2		0,053***	0,094***	-0,010***	-0,085***
NPF =3		0,059***	0,033***	0,015***	-0,048***
NPF =4		0,061***	0,013***	0,008***	-0,021***
RISQUE=2	Vin préventif et risqué	-0,024***	-0,010***	0,00004	0,010***
RISQUE=3	Sans effet	-0,066***	-0,038***	-0,004***	0,042***
RISQUE=4	Pas d'opinion	-0,105***	-0,070***	-0,022***	0,092***
RISQUE=5	Risqué	-0,127***	-0,101***	-0,048***	0,149***
SEXE=1	Homme	0,088***	0,157***	0,004	-0,162***
SOCF=1	Faible «sociabilité» du foyer	-0,071***	-0,095***	-0,017***	0,111***
TFOYBA=0	Pas de boisson alcoolisée	-0,244***	-0,129***	-0,104***	0,233***
TFOYBA=2	Autre b. alcoolisée quotidienne	-0,041***	-0,043***	0,001	0,043***
TFOYBA=3	Vin hebdomadaire	0,018***	-0,020***	-0,005***	0,024***
TFOYBA=4	Autre b. alcoolisée hebdomadaire	-0,022***	-0,060***	-0,035***	0,095***
TFOYBA=5	b. alcoolisée plus rarement	-0,035***	-0,085***	-0,043***	0,128***
VILLEENF=1	Ville d'enfance	-0,018***	-0,006***	-0,002***	0,007***
VALEUR=1	Insertion sociale	-0,025***	-0,008***	-0,0002	0,008***
VALEUR=3	Sécuritaire	-0,089***	-0,051***	-0,015***	0,065***
VALEUR=4	Dignité, respect de soi	-0,039***	-0,020***	-0,003***	0,023***