

Chapitre 3 : Modèle multinomial non ordonné

1) Présentation théorique

- Dans le cas du modèle logit ou probit ordonné, le caractère ordinal de la variable expliquée permettait d'introduire une variable latente et des seuils.
- Lorsque la variable est purement nominale, cette démarche n'est plus possible.

Hypothèses sous-jacentes :

L'individu est supposé évaluer les différentes alternatives qui s'offrent lui en se basant sur des critères (ex : prix, temps de trajets pour les modes de transport) et choisir l'alternative qui maximise son utilité.

Les différentes alternatives (choix) sont supposées mutuellement exclusives pour l'ensemble des individus.

L'utilité associée à chaque alternative est une variable aléatoire de la forme suivante :

$$U = V(X) + \varepsilon$$

V est la partie déterministe, fonction linéaire des différentes caractéristiques observées X des alternatives et/ou des individus dont les coefficients associés sont inconnus et à estimer

ε est la partie aléatoire (variables explicatives inobservées et erreur de mesure)

→ Différents types de modèles peuvent être utilisés pour l'analyse empirique des problèmes de choix discret (modèle multinomial, modèle conditionnel, modèle emboîté, etc.).

A) Présentation du modèle Logit multinomial (simple ou général) et du modèle Logit conditionnel

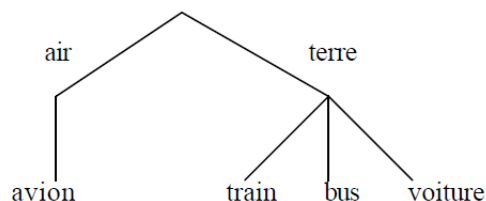
Ces deux types de modèles sont basés sur l'hypothèse d'indépendance des termes aléatoires des fonctions d'utilités.

Il ne tient pas compte d'éventuelles corrélations entre les alternatives.

Cette hypothèse est connue sous le sigle **IIA (Independence from Irrelevant Alternatives)** ou Indépendance par rapport aux choix non retenus.

Ces modèles ne prennent donc pas en compte la proximité de nature qui peut exister entre plusieurs choix offerts à l'individu. L'individu arbitre entre deux choix a et b indépendamment des autres choix qui lui sont offerts.

Or, dans le cas ci-dessous, cette hypothèse n'est pas vérifiée



→ Il faudra donc tester lors de l'estimation l'hypothèse d'IIA et utiliser si nécessaire des modèles emboîtés (*nested model*)

→ Modélisons le choix du mode de pêche

La variable dépendante y prend les valeurs 1, 2, 3, et 4 en fonction du mode de pêche :

- 1 : plage (*beach*)
- 2 : port (*pier*)
- 3 : bateau privé (*boat*)
- 4 : en bateau de groupe (*charter*)

• Les variables explicatives sont les suivantes :

- le revenu : fonction de l'individu, ne variant pas avec le mode de pêche
- le prix (P) et le taux de prise (C) variant avec le mode de pêche (alternative) et d'un individu à l'autre.

• 1182 observations.

A.1) Modèle Logit conditionnel

Le modèle Logit conditionnel a comme variables explicatives des caractéristiques liées aux différentes alternatives proposées.

Il ne contient aucune variable invariante d'un choix à l'autre.

Le modèle Logit conditionnel ne contient pas de termes constants.

Par ailleurs, les paramètres β du modèle sont indépendants du choix j.

La probabilité que l'individu i choisisse l'alternative l est définie de la manière suivante :

$$P_{il} = \text{Prob}[y_i = l | X] = \frac{\exp(\beta_p P_{il} + \beta_c C_{il})}{\sum_{j=1}^4 \exp(\beta_p P_{ij} + \beta_c C_{ij})}; \quad j = 1, \dots, 4$$

Où : P et C sont des variables explicatives dépendant des modalités (ici 4)

Interprétation des résultats :

Regressor	Type	Coefficient	CL
Price (P)	Specific	β_P	-0.021
Catch rate (C)	Specific	β_{CR}	0.953
- ln L			-1311

• $\beta_P < 0$

Une **hausse du prix** d'une alternative **diminue la probabilité de choisir cette alternative** et augmente la probabilité de choix d'autres alternatives.

• $\beta_{CR} > 0$

Une **hausse des prises** d'un mode de pêche **augmente la probabilité de choisir cette alternative** et diminue celle des autres alternatives.

A.2) Modèle Logit multinomial (au sens strict):

- Les valeurs des variables explicatives ne changent pas avec les alternatives mais changent avec les individus (exemple : revenu)
- Pour des raisons d'identification du modèle lors des estimations, il est nécessaire de modéliser (k-1) rapports de probabilités.
- Une alternative est donc toujours prise comme référence et les résultats obtenus sont à interpréter par rapport à cette référence.
- Le choix de cette alternative de référence peut être simple (non malade / différents types de maladies) mais si ce n'est pas le cas, on peut en choisir une arbitrairement.
- Cela n'a aucune incidence sur les calculs, seule l'interprétation des coefficients est différente.

La probabilité que l'individu i choisisse l'alternative l (autre que celle de référence) est :

$$p_{il} = \text{Prob}[y_i = l] = \frac{\exp(\alpha_l + \beta_l I_i)}{1 + \sum_{j=1}^3 \exp(\alpha_j + \beta_j I_i)}; \quad j = 1, \dots, 3$$

La probabilité que l'individu choisisse l'alternative de référence est :

$$p_{i4} = \text{Prob}[y_i = 4] = \frac{1}{1 + \sum_{j=1}^3 \exp(\alpha_j + \beta_j I_i)}; \quad j = 1, \dots, 3$$

Où : I mesure le revenu.

➔ Dans ce cas, les coefficients des variables explicatives changent avec les alternatives.

Regressor	Type	Coefficient	Model type	
			CL	MNL
Price (P)	Specific	β_P	-0.021	-
Catch rate (C)	Specific	β_{CR}	0.953	-
Intercept	Invariant	$\alpha_1 : Beach$	-	0.0
		$\alpha_2 : Pier$	-	0.814
		$\alpha_3 : Private$	-	0.739
		$\alpha_4 : Charter$	-	1.341
			-	
Income (I)	Invariant	$\beta_{I1} : Beach$	-	0.0
		$\beta_{I2} : Pier$	-	-0.143
		$\beta_{I3} : Private$	-	0.092
		$\beta_{I4} : Charter$	-	-0.032
			-	
- ln L			-1311	-1477

- ➔ Si le revenu de la personne augmente, la probabilité de choisir le port par rapport à la plage diminue. Il en est de même pour la pêche en groupe.
- ➔ Inversement, la probabilité de choisir la pêche en bateau privé par rapport à la plage augmente.

→ Il existe un **flottement terminologique** dans la littérature entre le Logit multinomial et le Logit conditionnel où le mot multinomial peut être utilisé pour un modèle multinomial simple ou un modèle combinant les deux types de caractéristiques, idem pour le terme de Logit conditionnel qui peut être utilisé pour un modèle conditionnel ou un modèle combinant les deux.

→ De plus, ce sont souvent les mêmes procédures informatiques qui sont utilisées pour estimer indifféremment un Logit multinomial ou un Logit conditionnel (ex : mlogit) à condition que les données soient configurées de manière adéquate.

A.3) Modèle logit multinomial général

Un modèle plus général regroupe donc les deux modèles précédents.

Si on reprend l'exemple précédent :

$$p_{il} = \text{Prob}[y_i = l] = \frac{\exp(\beta_P P_{il} + \beta_C C_{il} + \alpha_l + \beta_I I_i)}{1 + \sum_{j=1}^3 \exp(\beta_P P_{ij} + \beta_C C_{ij} + \alpha_j + \beta_I I_i)}; \quad j = 1, \dots, 3$$

$$p_{i4} = \text{Prob}[y_i = 4] = \frac{1}{1 + \sum_{j=1}^3 \exp(\beta_P P_{ij} + \beta_C C_{ij} + \alpha_j + \beta_I I_i)}; \quad j = 1, \dots, 3$$

Regressor	Type	Coefficient	Model type		
			CL	MNL	MNL G
Price (P)	Specific	β_P	-0.021	-	-0.025
Catch rate (C)	Specific	β_{CR}	0.953	-	0.358
Intercept	Invariant	$\alpha_1 : Beach$	-	0.0	0.0
		$\alpha_2 : Pier$	-	0.814	0.778
		$\alpha_3 : Private$	-	0.739	0.527
		$\alpha_4 : Charter$	-	1.341	1.694
Income (I)	Invariant	$\beta_{I1} : Beach$	-	0.0	0.0
		$\beta_{I2} : Pier$	-	-0.143	-0.128
		$\beta_{I3} : Private$	-	0.092	0.089
		$\beta_{I4} : Charter$	-	-0.032	-0.033
-ln L			-1311	-1477	-1215

B) Méthode d'estimation

La méthode d'estimation des différents paramètres du modèle est **toujours basée sur la méthode du maximum de vraisemblance**.

La fonction de vraisemblance pour un échantillon de n observations indépendantes est :

$$L_n = \prod_{i=1}^n \prod_{j=1}^4 p_{ij}^{y_{ij}}$$

La fonction de vraisemblance est alors :

$$\text{Log } L_n = \sum_{i=1}^n \sum_{j=1}^4 y_{ij} \ln p_{ij}$$

Avec $p_{ij} = F(X_i, \beta)$

Où : F est la répartition du modèle Logit

Les conditions du premier ordre permettent de terminer les coefficients à estimer :

$$\frac{\partial \text{Log } L_n}{\partial \beta} = 0$$

Ratio de vraisemblance :

Ce test doit être également utilisé pour vérifier l'intérêt du modèle. Il est nécessaire dans ce cas de calculer le modèle contraint (sans les variables explicatives).

Qualité de l'ajustement du modèle aux données

La qualité de l'estimation peut également se mesurer par le R^2 Mac Fadden (la plus part du temps indiqué dans les résultats de l'estimation)

Qualité de prévision du modèle estimé :

Il est nécessaire également de mesurer la qualité de la prévision pour chacune des probabilités associées à chaque alternative.

C) Effet marginaux et odd-ratio

Effets marginaux :

On peut mesurer l'effet de chaque variable explicative sur les différents choix en dérivant les probabilités de choix par rapport à la variable.

→ Calcul de la dérivée partielle de chaque probabilité par rapport à la variable. On mesure alors l'effet de l'augmentation *d'une unité* de la variable sur chaque choix.

Cas du modèle conditionnel sur la pêche :

L'effet marginal d'une variable explicative se calcule de la manière suivante :

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial p_{ij}}{\partial x_{ipj}}$$

→ Dans le cas d'un modèle multinomial stricte, la formule est la même mais l'interprétation se fait par rapport à la probabilité de référence.

Odd ratio

Variable quantitative :

Les individus dont la variable explicative augmente d'un point ont x fois plus de chance de choisir l'alternative j (/ celle de référence dans le cas d'un modèle multinomial) par rapport à ceux dont la valeur de la variable explicative n'a pas changé.

Variable binaire (passage de 0 à 1)

Les individus ayant la caractéristique ont x fois plus de chance de choisir l'alternative j (/ celle de référence dans le cas d'un modèle multinomial) par rapport à ceux qui n'ont pas cette caractéristique.

2) Application sous R

2.1) Utilisation de la fonction `mlogit`

Cette fonction permet d'estimer les deux types de modèles multinomiaux (simples et généraux) et conditionnels.

➔ Cette fonction nécessite néanmoins une préparation de la base qui diffèrent selon le format initial de la base de données

En effet, les bases de données peuvent avoir deux formats.

- Un format appelé « wide shape » : il y a une ligne pour chaque individu où son choix est indiqué (ex: 200 individus)
- Un format appelé « long shape » : il y a dans ce cas une ligne pour chaque alternative pour chaque situation de choix (ex : 4 choix pour 200 individus ➔ 800 lignes)

Exemple : Reprenons l'exemple du choix du mode de pêche

Cette base est disponible dans la librairie `mlogit` sous format « wide shape »

```
library(mlogit)
data("Fishing")
head(Fishing,3)
```

Prix des différentes pratiques de pêche
(prix et alternative séparés par un .)

	mode	price.beach	price.pier	price.boat	price.charter	catch.beach
1	charter	157.930	157.930	157.930	182.930	0.0678
2	charter	15.114	15.114	10.534	34.534	0.1049
3	boat	161.874	161.874	24.334	59.334	0.5333
		catch.pier	catch.boat	catch.charter	income	
1		0.0503	0.2601	0.5391	7083.332	
2		0.0451	0.1574	0.4671	1250.000	
3		0.4522	0.2413	1.0266	3750.000	

Revenu pour
chaque individu
invariant/alternative

Capture pour les différentes pratiques de pêche

Inconvénient : ce format oblige à créer beaucoup de colonnes (il y a autant de colonnes de prix que d'alternatives, idem pour les prises)

Format « long SHAPE »

```
library(AER)
data("TravelMode")
head(TravelMode,6)
```

	individual	mode	choice	wait	vcost	travel	gcost	income	size
1	1	air	no	69	59	100	70	35	1
2	1	train	no	34	31	372	71	35	1
3	1	bus	no	35	25	417	70	35	1
4	1	car	yes	0	10	180	30	35	1
5	2	air	no	64	58	68	68	30	2
6	2	train	no	44	31	354	84	30	2

→ Base : 840 observations : 210*4

→ Car 4 modes de transport (air, train, bus et voitures) (variable mode) et 210 individus

	individual	mode	choice	wait	vcost	travel	gcost	income	size
1	1	air	no	69	59	100	70	35	1
2	1	train	no	34	31	372	71	35	1
3	1	bus	no	35	25	417	70	35	1
4	1	car	yes	0	10	180	30	35	1
5	2	air	no	64	58	68	68	30	2
6	2	train	no	44	31	354	84	30	2

→ 4 variables spécifiques aux alternatives :

wait : temps d'attente au terminal (0 pour la voiture)

vcost : coût lié au véhicule

travel : temps du trajet

gcost : coût global.

→ Deux variables spécifiques aux individus : le revenu (income) et la taille du ménage (size)

→ La variable choice indique si l'alternative a été choisie par l'individu

Transformation d'un format wide en un format long :

```
Fish<-mlogit.data(Fishing, shape= "wide",varying=2:9,choice= "mode")
head(Fish,8)
```

varying =2:9 : indique que les colonnes 2 à 9 correspondent à des variables qui changent avec les alternatives.

choice = « mode » : indique que c'est la variable mode qui sera la variable à expliquer

	<i>mode</i>	<i>income</i>	<i>alt</i>	<i>price</i>	<i>catch</i>	<i>chid</i>
1.beach	FALSE	7083.332	beach	157.930	0.0678	1
1.boat	FALSE	7083.332	boat	157.930	0.2601	1
1.charter	TRUE	7083.332	charter	182.930	0.5391	1
1.pier	FALSE	7083.332	pier	157.930	0.0503	1
2.beach	FALSE	1250.000	beach	15.114	0.1049	2
2.boat	FALSE	1250.000	boat	10.534	0.1574	2
2.charter	TRUE	1250.000	charter	34.534	0.4671	2
2.pier	FALSE	1250.000	pier	15.114	0.0451	2

....

➔ Variable chid (identifiant de l'individu) et variable alt (identifiant du choix) : créés lors de ce changement de format

Importation d'un format long

```
Transport<-mlogit.data(TravelMode, choice="choice",shape="long",alt.var="mode")
head(Transport,8)
```

	<i>individual</i>	<i>mode</i>	<i>choice</i>	<i>wait</i>	<i>vcost</i>	<i>travel</i>	<i>gcost</i>	<i>income</i>	<i>size</i>
1.air	1	air	FALSE	69	59	100	70	35	1
1.train	1	train	FALSE	34	31	372	71	35	1
1.bus	1	bus	FALSE	35	25	417	70	35	1
1.car	1	car	TRUE	0	10	180	30	35	1
2.air	2	air	FALSE	64	58	68	68	30	2
2.train	2	train	FALSE	44	31	354	84	30	2
2.bus	2	bus	FALSE	53	25	399	85	30	2
2.car	2	car	TRUE	0	11	255	50	30	2

* **Estimation du modèle conditionnel (pas besoin de choix de référence) (ex : choix du mode de pêche)**

```
ml.Fish<-mlogit(mode~price+catch|0,data=Fish)
summary(ml.Fish)
```

Frequencies of alternatives:

beach	boat	charter	pier
0.11337	0.35364	0.38240	0.15059

.....

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t)
price	-0.0204765	0.0012231	-16.742	< 2.2e-16 ***
catch	0.9530982	0.0894134	10.659	< 2.2e-16 ***

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

Log-Likelihood: -1312

Rappel :

- $\beta_p < 0$: indique qu'une hausse du prix d'une alternative **diminue de manière significative la probabilité de choisir cette alternative** et augmente la probabilité de choix d'autres alternatives.

- $\beta_{CR} > 0$: indique qu'une hausse des prises d'un mode de pêche **augmente la probabilité de manière significative de choisir cette alternative** et diminue celle des autres alternatives.

Remarque : dans le cas du modèle conditionnel, il faut calculer le R^2 de Mac Fadden ainsi que le ratio de vraisemblance (ce qui ne sera pas le cas des modèles multinomiaux estimés par la fonction mlogit)

Calcul de la qualité du modèle : $R^2_{Mc_Fadden}$

```
ml.Fish0<-mlogit(mode~1,data=Fish)
R2_Mc_Fadden<-1-(logLik(ml.Fish)/logLik(ml.Fish0))
R2_Mc_Fadden
'log Lik.' 0.1240171 (df=2)
```

```
library(lmtest)
lrtest(ml.Fish0,ml.Fish)
```

Likelihood ratio test

Model 1: mode ~ 1

Model 2: mode ~ price + catch | 0

```
#Df LogLik Df Chisq Pr(>Chisq)
1 3 -1497.7
2 2 -1312.0 -1 371.49 < 2.2e-16 ***
```

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

➔ Au seuil de risque de 1% (et donc de 5%), intérêt d'estimer le modèle avec les variables explicatives price et catch

* Modèle multinomial simple

```
ml.Fish2<-mlogit(mode~0|income,data=Fish)
summary(ml.Fish2)
```

	Estimate	Std. Error	t-value	Pr(> t)
boat:(intercept)	7.3892e-01	1.9673e-01	3.7560	0.0001727 ***
charter:(intercept)	1.3413	1.9452e-01	6.8955	5.367e-12 ***
pier:(intercept)	8.1415e-01	2.2863e-01	3.5610	0.0003695 ***
boat:income	9.1906e-05	4.0664e-05	2.2602	0.0238116 *
charter:income	-3.1640e-05	4.1846e-05	-0.7561	0.4495908
pier:income	-1.4340e-04	5.3288e-05	-2.6911	0.0071223 **

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

Log-Likelihood: -1477.2

McFadden R^2: 0.013736

Likelihood ratio test : chisq = 41.145 (p.value = 6.0931e-09)

Interprétation des résultats :

- ➔ Cette fonction permet d'obtenir directement le R2 de Mac Fadden et le test du ratio de vraisemblance
- ➔ Le fait de n'introduire que les variables liées aux individus diminue la qualité du modèle / modèle logit conditionnel
- ➔ Si le revenu de la personne augmente, la probabilité de choisir le port par rapport à la plage diminue.
- ➔ Inversement, la probabilité de choisir la pêche en bateau privé par rapport à la plage augmente si le revenu augmente. Ces impacts sont significatifs.
- ➔ Il n'y a pas d'effet significatif d'une variation du revenu sur la probabilité de choisir la pêche en bateau en groupe par rapport à la plage

* Modèle multinomial général

```
ml.Fish3<-mlogit(mode~price+catch|income,data=Fish)
summary(ml.Fish3)
```

Frequencies of alternatives:

<i>beach</i>	<i>boat</i>	<i>charter</i>	<i>pier</i>
0.11337	0.35364	0.38240	0.15059

	<i>Estimate</i>	<i>Std. Error</i>	<i>t-value</i>	<i>Pr(> t)</i>
<i>boat:(intercept)</i>	5.2728e-01	2.2279e-01	2.3667	0.0179485 *
<i>charter:(intercept)</i>	1.6944	2.2405e-01	7.5624	3.952e-14 ***
<i>pier:(intercept)</i>	7.7796e-01	2.2049e-01	3.5283	0.0004183 ***
<i>price</i>	-2.5117e-02	1.7317e-03	-14.5042	< 2.2e-16 ***
<i>catch</i>	3.5778e-01	1.0977e-01	3.2593	0.0011170 **
<i>boat:income</i>	8.9440e-05	5.0067e-05	1.7864	0.0740345 .
<i>charter:income</i>	-3.3292e-05	5.0341e-05	-0.6613	0.5084031
<i>pier:income</i>	-1.2758e-04	5.0640e-05	-2.5193	0.0117582 *

Log-Likelihood: -1215.1
 McFadden R^2: 0.18868
 Likelihood ratio test : $\chi^2 = 565.17$ ($p.value = < 2.22e-16$)

→ C'est le meilleur modèle des 3 modèles estimés.

→ Il est intéressant de remarquer que l'effet du revenu sur la probabilité de choisir le bateau privé par rapport à la plage lorsque le revenu augmente devient moins significatif.

Au seuil de risque de 5 %, seule une augmentation du revenu fait diminuer de manière significative le fait de choisir le port par rapport à la plage.

Remarque :

→ L'alternative prise comme référence est l'alternative « beach ».

→ Si on veut changer la référence dans le modèle multinomial combiné, il suffit d'introduire l'option **reflevel**

ml.Fish4<-mlogit(mode~price+catch|income,data=Fish, reflevel="boat")
summary(ml.Fish4)

....

	Estimate	Std. Error	t-value	Pr(> t)
beach:(intercept)	-5.2728e-01	2.2279e-01	-2.3667	0.017948 *
charter:(intercept)	1.1671	1.5905e-01	7.3380	2.169e-13 ***
pier:(intercept)	2.5068e-01	2.0394e-01	1.2292	0.219000
price	-2.5117e-02	1.7317e-03	-14.5042	< 2.2e-16 ***
catch	3.5778e-01	1.0977e-01	3.2593	0.001117 **
beach:income	-8.9440e-05	5.0067e-05	-1.7864	0.074035 .
charter:income	-1.2273e-04	2.8631e-05	-4.2867	1.813e-05 ***
pier:income	-2.1702e-04	5.0058e-05	-4.3353	1.456e-05 ***

Log-Likelihood: -1215.1
 McFadden R^2: 0.18868
 Likelihood ratio test : $\chi^2 = 565.17$ ($p.value = < 2.22e-16$)

Reprenons l'estimation `ml.Fish3` (avec "beach" comme modalité de référence)

```
ml.Fish3<-mlogit(mode~price+catch|income,data=Fish)
summary(ml.Fish3)
```

→ **Calcul des effets marginaux pour la variable explicative Income**

Exemple : effet du revenu sur les différentes probabilités pour chaque individu

```
effects(ml.Fish3, covariate = "income", data = Fish)
```

	<i>beach</i>	<i>boat</i>	<i>charter</i>	<i>pier</i>
[1,]	-1.614352e-06	3.265371e-05	-1.567098e-05	-1.536841e-05
[2,]	2.141171e-06	2.708020e-05	-6.260531e-06	-2.296084e-05
[3,]	-1.598014e-07	3.082468e-05	-2.930347e-05	-1.361417e-06
[4,]	9.016848e-06	2.199539e-05	1.643405e-06	-3.265563e-05
[5,]	-7.832006e-07	3.195502e-05	-2.477757e-05	-6.394239e-06
...				

Calcul des effets marginaux pour le revenu mais au niveau moyen de l'échantillon

→ Créer un tableau contenant les valeurs moyennes des variables explicatives pour chaque alternative

```
(moy <- with(Fish, data.frame(price = tapply(price, index(Fish)$alt, mean),
catch = tapply(catch, index(Fish)$alt, mean), income = mean(income))))
```

	<i>price</i>	<i>catch</i>	<i>Income</i>
<i>beach</i>	103.42201	0.2410113	4099.337
<i>boat</i>	55.25657	0.1712146	4099.337
<i>charter</i>	84.37924	0.6293679	4099.337
<i>pier</i>	103.42201	0.1622237	4099.337

`effects(ml.Fish3, covariate = "income", data = moy)`

<i>beach</i>	<i>boat</i>	<i>charter</i>	<i>pier</i>
-7.214174e-07	3.176131e-05	-2.173391e-05	-9.305980e-06

→ Une augmentation du revenu d'une unité (/niveau moyen de l'échantillon) conduira par exemple à une baisse de la probabilité de choisir le mode charter de $-2.173e-05$

Calcul des effets marginaux au niveau moyen de l'échantillon (/ élasticités)

`effects(ml.Fish3, covariate = "income", type="rr", data = moy)`

<i>beach</i>	<i>boat</i>	<i>charter</i>	<i>pier</i>
-0.05634297	0.31030091	-0.19281695	-0.57932471

Calcul des effets marginaux pour le prix au niveau moyen de l'échantillon

`effects(ml.Fish3, covariate = "price", data = moy)`

	<i>beach</i>	<i>boat</i>	<i>charter</i>	<i>pier</i>
<i>beach</i>	-1.249124e-03	0.0005531588	0.0006091541	8.681094e-05
<i>boat</i>	5.531588e-04	-0.0061167595	0.0048696270	6.939737e-04
<i>charter</i>	6.091542e-04	0.0048696271	-0.0062430047	7.642235e-04
<i>pier</i>	8.681094e-05	0.0006939736	0.0007642234	-1.545008e-03

→ Une augmentation d'une unité du prix (1 \$) / plage (/au niveau moyen) aura pour impact une baisse de la probabilité de choisir la plage et mais aura également un impact positif sur les autres modes de pêches

→ Calcul des odd-ratios par exemple pour le modèle multinomial simple

```
ml.Fish2<-mlogit(mode~0|income,data=Fish)
exp(coef(ml.Fish2))
```

```
boat:(intercept) charter:(intercept) pier:(intercept)
2.0936747      3.8239787      2.2572568
```

```
boat:income charter:income pier:income
1.0000919    0.9999684    0.9998566
```

→ Les individus dont le revenu augmente d'une unité ont 1/0,99985 (1,000143) moins de chance de choisir le port / plage que ceux dont le revenu n'a pas changé

2.2) Utilisation de la fonction vglm

→ Cette fonction permet d'estimer les modèles multinomiaux simples (pour les modèles conditionnels et multinomiaux généraux utiliser la fonction mlogit).

L'avantage de cette fonction (vglm) est qu'elle ne nécessite pas de transformation de la base de données dans le cas d'un modèle multinomial simple.

De plus, elle permet de représenter graphiquement l'effet d'une variable explicative quantitative sur les différentes probabilités des alternatives j et de calculer pour chaque individu la probabilité associée à chaque alternative

```
library(mlogit)
data("Fishing")
```

```
library(VGAM)
Fit<-vglm(mode~income,multinomial(refLevel=1),data=Fishing)
summary(Fit)
```

Coefficients:

	Estimate	Std. Error	z-value	Pr(> z)
(Intercept):1	8.142e-01	2.286e-01	3.561	0.000369 ***
(Intercept):2	7.389e-01	1.967e-01	3.756	0.000173 ***
(Intercept):3	1.341e+00	1.945e-01	6.896	5.37e-12 ***
income:1	-1.434e-04	5.329e-05	-2.691	0.007121 **
income:2	9.191e-05	4.066e-05	2.260	0.023814 *
income:3	-3.164e-05	4.185e-05	-0.756	0.449594

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

Residual deviance: 2954.301 on 3540 degrees of freedom

Log-likelihood: -1477.151 on 3540 degrees of freedom

str(Fishing)

\$ price.beach : num 157.9 15.1 161.9 15.1 106.9 ...

\$ price.pier : num 157.9 15.1 161.9 15.1 106.9 ...

\$ price.boat : num 157.9 10.5 24.3 55.9 41.5 ...

\$ price.charter: num 182.9 34.5 59.3 84.9 71 ...

➔ Par conséquent l'alternative de référence est « beach »

Rappel : avec la fonction mlogit (lorsque “beach” est pris comme référence par défaut)

	Estimate	Std. Error	t-value	Pr(> t)
boat:(intercept)	7.3892e-01	1.9673e-01	3.7560	0.0001727 ***
charter:(intercept)	1.3413	1.9452e-01	6.8955	5.367e-12 ***
pier:(intercept)	8.1415e-01	2.2863e-01	3.5610	0.0003695 ***
boat:income	9.1906e-05	4.0664e-05	2.2602	0.0238116 *
charter:income	-3.1640e-05	4.1846e-05	-0.7561	0.4495908
pier:income	-1.4340e-04	5.3288e-05	-2.6911	0.0071223 **

Intérêt du modèle :

Fit0<-vglm(mode~1,multinomial(refLevel=1),data=Fishing)

1-pchisq(deviance(Fit0)-deviance(Fit) , df = df.residual(Fit0) - df.residual(Fit))

[1] 6.093087e-09

Calcul du R2 de Mac Fadden :

R2_Mc_Fadden<-1-(deviance(Fit)/deviance(Fit0))

R2_Mc_Fadden

[1] 0.01373575

Calcul des probabilités pour chaque alternative pour chaque individu

(fitted(Fit))

	<i>beach</i>	<i>pier</i>	<i>boat</i>	<i>charter</i>
1	0.11250922	0.09196564	0.4516733	0.3438518
2	0.11221981	0.21173938	0.2635553	0.4124855
3	0.11534399	0.15206503	0.3408657	0.3917253
4	0.11376764	0.19048097	0.2884583	0.4072931
5	0.11536587	0.13496232	0.3680676	0.3816042
6	0.11536587	0.13496232	0.3680676	0.3816042
7	0.10843710	0.06979378	0.5073859	0.3143832
...				

→ Pour lire plus facilement les coefficients du modèle :

head(coef(Fit,matrix=TRUE))

	$\log(\mu_{[2]}/\mu_{[1]})$	$\log(\mu_{[3]}/\mu_{[1]})$	$\log(\mu_{[4]}/\mu_{[1]})$
(Intercept)	0.8141502697	7.389208e-01	1.341291e+00
income	-0.0001434029	9.190636e-05	-3.163988e-05

→ Calcul des odd-ratios:

exp(coef(Fit))

(Intercept):1	(Intercept):2	(Intercept):3	income:1	income:2	income:3
2.2572568	2.0936747	3.8239787	0.9998566	1.0000919	0.9999684

Calcul des effets marginaux pour chaque personne :

margeff(Fit)

, , 1

	<i>beach</i>	<i>pier</i>	<i>boat</i>	<i>charter</i>
(Intercept)	-9.786391e-02	-0.0051206362	-5.912828e-02	1.621128e-01
income	-1.962620e-06	-0.0000147924	3.363263e-05	-1.687761e-05

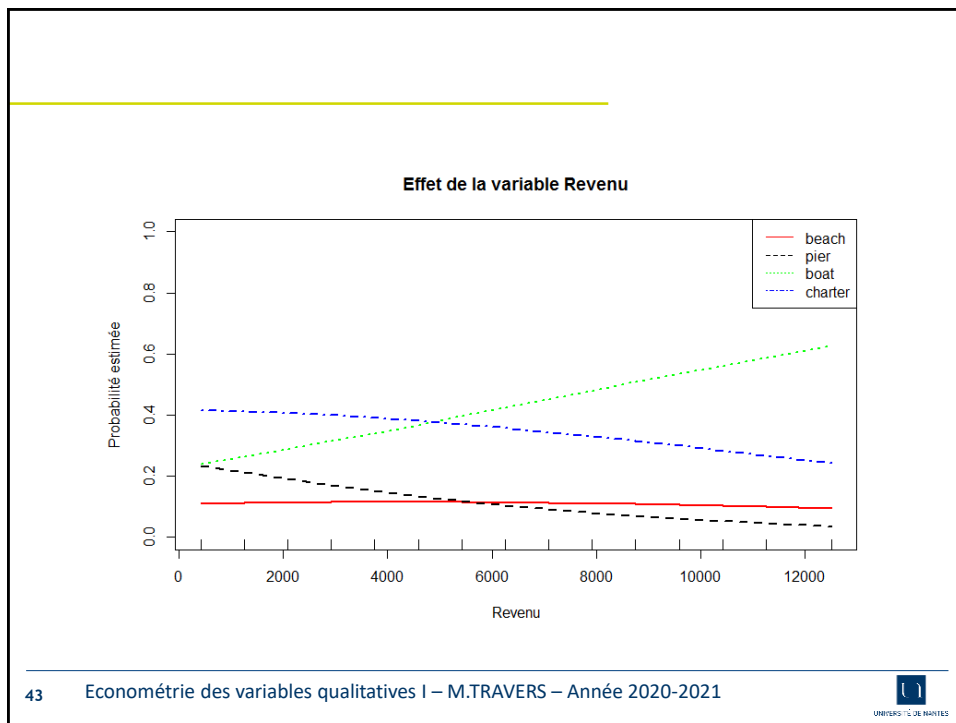
, , 2

	<i>beach</i>	<i>pier</i>	<i>boat</i>	<i>charter</i>
(Intercept)	-1.032868e-01	-2.249671e-02	-4.782914e-02	1.736127e-01
income	2.153793e-06	-2.630021e-05	2.928073e-05	-5.134311e-06

...

Représentation graphique de l'effet revenu sur les probabilités de choix :

```
mycol <- c("red", "green", "blue")
ooo <- with(Fishing, order(income))
with(Fishing, matplot(income[ooo], fitted(Fit)[ooo,], ylim = c(0,1),
  xlab = "Revenu", ylab = "Probabilité estimée",
  main = " Effet de la variable Revenu ", type = "l", lwd = 2, col = c(mycol[1],
    "black", mycol[-1])))
with(Fishing, rug(income))
legend("topright", col = c(mycol[1], "black", mycol[-1]), lty=1:4
, legend=colnames(Fit@y))
```



3) Extensions du modèle multinomial

3.1) Prise en compte de l'hétéroscédasticité des erreurs

Hypothèse d'homoscédasticité des erreurs

```
ml.Fish3<-mlogit(mode~price+catch|income,data=Fish)
```

#Hypothèse d'hétéroscédasticité des erreurs

```
ml.Fish3het<-mlogit(mode~price+catch|income, heterosc = TRUE, data=Fish)
```

Différence entre les deux modèles

```
lrtest(ml.Fish3,ml.Fish3het)
```

Likelihood ratio test

Model 1: mode ~ price + catch | income

Model 2: mode ~ price + catch | income

# Df	LogLik	Df	Chisq	Pr(>Chisq)
1	8	-1215.1		
2	11	-1218.2	3	6.0495 0.1092

➔ Au seuil de risque de 10%, il n'y pas de différence entre les deux modèles

3.2) Utilisation du modèle de type emboîté ou « nested model » :

- ➔ L'expression des probabilités du modèle multinomial est liée au fait que les termes aléatoires sont indépendants et identiquement distribués : indépendance des états non pertinents (IIA)
- ➔ Le rapport des probabilités associées au choix entre deux modalités est indépendant des autres modalités. Ajouter ou enlever une tierce modalité ne change pas le rapport entre ces deux probabilités.
- ➔ Cette propriété peut être acceptée si les modalités sont peu comparables mais elle devient discutable lorsque le modèle multinomial (Logit) est utilisé pour modéliser des choix très similaires.
- ➔ Il faut donc tester cette hypothèse sachant que l'hypothèse H_0 correspond au fait d'accepter l'hypothèse IIA.

Pour cela, on utilise la fonction `mlogit` sous R.

Vérification de l'hypothèse d'absence de corrélation sous R

Pour tester l'absence de « nœuds », il y a deux tests possibles :

Premier test possible : `lrtest(modele nested)`

2^{ème} test possible : `scoretest(modele multinomial, nests = list(nom de la branche1 = c(' ', ' ', ' '), nom de la branche 2 = c(' ', ' ', ' '),`)

Si la p-value est inférieure à 0,05 (ou 0,1), l'hypothèse IIA est refusée.

Reprenons l'exemple du transport :

- ➔ Il faut tout d'abord faire des hypothèses sur les similarités entre les choix (branche)
- ➔ Il est possible de supposer que le transport par air est différent des autres modes de transport. À l'inverse, on peut supposer que la voiture, le bus et le train sont des moyens similaires

Application du premier test :

```
library(lmtest)
library(mlogit)
```

```
data("TravelMode", package= "AER")
Transport<-mlogit.data(TravelMode, choice="choice",shape="long",alt.var="mode")
```

```
nl.Transport<-
mlogit(choice~wait+vcost+travel|income+size,data=Transport,reflevel="train",nests=lis
t(air="air",terre=c("bus","car","train")),unscaled=TRUE)
```

Remarque :

Si une branche de l'arbre ne comporte qu'une modalité (arbre dégénéré), il est nécessaire de mettre dans l'option nests : unscaled=TRUE

lrtest(nl.Transport)

Likelihood ratio test

Model 1: choice ~ wait + vcost + travel | income + size

Model 2: choice ~ wait + vcost + travel | income + size

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
1	14	-153.74			
2	12	-172.47	-2	37.453	7.365e-09 ***

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

Deuxième type de test

```
ml.Transport<-
mlogit(choice~wait+vcost+travel|income+size,data=Transport,reflevel="train")
scoretest(ml.Transport,
nests=list(air="air",terre=c("bus","car","train")),unscaled=TRUE)
```

score test
data: nests = list(air = c('air'), terre = c('bus','car','train'))
chisq = 26.7648, df = 2, p-value = 1.542e-06
alternative hypothesis: nested model

Conclusion :

Pour les deux tests, l'hypothèse IIA est refusée au seuil de 5 % pour les hypothèses émises

4) Application sous Stata

A) Comment estimer un modèle multinomial général sous Stata

Soit l'exemple suivant : choix du type d'automobile (américaine, japonaise ou européenne) de 295 consommateurs

Variables à expliquer

choice : 1 pour indiquer le type d'automobile choisi (américaine, japonaise, européenne), 0 dans le cas contraire

car : alternative (origine de la voiture)

Variables explicatives :

dealer : nombre de concessionnaires de chaque type de voiture dans la ville du consommateur

sex : genre du consommateur

income : revenu en milliers de dollars

size: taille du ménage du consommateur

La base est déjà en format long (3 choix possibles * 295 individus)

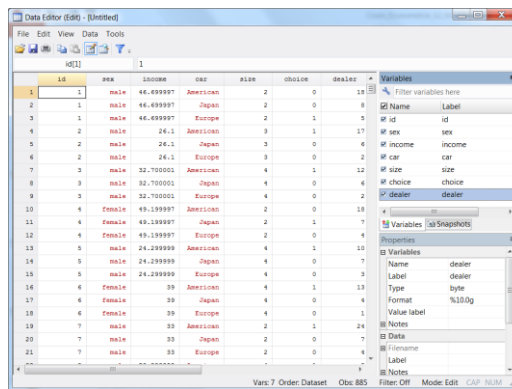
Importation d'une base excel sous Stat

clear

import excel "C:\Users\travers-

m\Desktop\Cours_2020_2021\Econometrie_variables_qualitatives_M1_EKAP_M2_CO

DEME\Cours\Voiture.xls", sheet("Sheet1") firstrow



51 Econométrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



describe

Contains data

obs: 885

vars: 7

size: 23,895

variable name	storage type	display format	value label	variable label
id	int	%10.0g		id
sex	str6	%9s		sex
income	double	%10.0g		income
car	str8	%9s		car
size	byte	%10.0g		size
choice	byte	%10.0g		choice
dealer	byte	%10.0g		dealer

52 Econométrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



list id car choice dealer sex income size in 1/12, sepy(id)

	id	car	choice	dealer	sex	income	size
1.	1	American	0	18	male	46.699997	2
2.	1	Japan	0	8	male	46.699997	2
3.	1	Europe	1	5	male	46.699997	2
4.	2	American	1	17	male	26.1	3
5.	2	Japan	0	6	male	26.1	3
6.	2	Europe	0	2	male	26.1	3
7.	3	American	1	12	male	32.700001	4
8.	3	Japan	0	6	male	32.700001	4
9.	3	Europe	0	2	male	32.700001	4
10.	4	American	0	18	female	49.199997	2
11.	4	Japan	1	7	female	49.199997	2
12.	4	Europe	0	4	female	49.199997	2

codebook sex

```

type: string (str6)
unique values: 2 missing "": 0/885
tabulation: Freq. Value
              237 "female"
              648 "male"

```

codebook dealer

```

type: numeric (byte)
range: [1,24] units: 1
unique values: 22 missing .: 0/885
mean: 9.99322
std. dev: 7.14538
percentiles: 10% 25% 50% 75% 90%
              3 4 8 17 22

```

codebook car

```
type: string (str8)
unique values: 3          missing "": 0/885
tabulation: Freq. Value
             295 "American"
             295 "Europe"
             295 "Japan"
```

Estimation du modèle :

xi:asclogit choice dealer, case(id) alternatives(car) casevars(i.sex income size)

Remarque : l'alternative de référence est par défaut l'alternative ayant la plus haute fréquence (ici American, voir slide 69)

Si on souhaite une alternative de référence différente, il suffit de rajouter **basealternative(nom de l'alternative)** à la fin de la commande ci-dessus

Alternative-specific conditional logit	Number of obs	=	885
Case variable: id	Number of cases	=	295
Alternative variable: car	Alts per case: min	=	3
	avg	=	3.0
	max	=	3
Log likelihood = -244.516	Wald chi2(7)	=	121.43
	Prob > chi2	=	0.0000

choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
car						
dealer	.1424601	.0300018	4.75	0.000	.0836577	.2012625
American	(base alternative)					
Europe						
_Isex_2	.4446935	.4107999	1.08	0.279	-.3604595	1.249847
income	.0271243	.0126302	2.15	0.032	.0023695	.051879
size	-.3263475	.1459232	-2.24	0.025	-.6123516	-.0403434
Japan						
_Isex_2	-.3517742	.3117905	-1.13	0.259	-.9628723	.2593238
income	.0510144	.011163	4.57	0.000	.0291353	.0728935
size	-.5896642	.1416146	-4.16	0.000	-.8672237	-.3121047

Avec _Isex_2 for sex==male

estat ic

Akaike's information criterion and Bayesian information criterion

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	885	.	-244.516	7	503.032	536.5311

Note: N=Obs used in calculating BIC; see [\[R\] BIC note](#).

Pour obtenir les odd-ratios associés à chacune des variables explicatives du modèle précédent :

asclogit, or

choice	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
car						
dealer	1.153107	.0345953	4.75	0.000	1.087257	1.222946
American	(base alternative)					
Europe						
_Isex_2	1.560012	.6408528	1.08	0.279	.6973558	3.489808
income	1.027496	.0129775	2.15	0.032	1.002372	1.053248
size	.7215544	.1052915	-2.24	0.025	.5420746	.9604596
Japan						
_Isex_2	.7034389	.2193255	-1.13	0.259	.3817947	1.296053
income	1.052338	.0117473	4.57	0.000	1.029564	1.075616
size	.5545135	.0785272	-4.16	0.000	.4201163	.7319049

Calcul des effets marginaux pour chacune des variables explicatives pour chaque type d'automobile (au niveau moyen de l'échantillon avec comme dummy la variable sex):

estat mfx

Pr(choice = American|1 selected) = .67638735

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]		X
dealer							
American	.031183	.006252	4.99	0.000	.01893	.043436	18.976
Europe	-.01317	.002981	-4.42	0.000	-.019014	-.007327	3.461
Japan	-.018012	.004008	-4.49	0.000	-.025868	-.010157	7.5424
casevars							
_Isex_2	.003365	.06077	0.06	0.956	-.115742	.122473	.7322
income	-.008958	.002273	-3.94	0.000	-.013413	-.004502	42.097
size	.104726	.023436	4.47	0.000	.058792	.15066	2.6237

Pr(choice = Europe|1 selected) = .13668252

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
dealer								
American	-.01317	.002981	-4.42	0.000	-.019014	-.007327		18.976
Europe	.01681	.00365	4.60	0.000	.009656	.023965		3.461
Japan	-.00364	.000835	-4.36	0.000	-.005276	-.002003		7.5424
casevars								
_Isex_2	.061462	.047366	1.30	0.194	-.031374	.154298		.7322
income	.001897	.001394	1.36	0.173	-.000834	.004629		42.097
size	-.023443	.016657	-1.41	0.159	-.056091	.009205		2.6237

Pr(choice = Japan|1 selected) = .18693013

variable	dp/dx	Std. Err.	z	P> z	[95% C.I.]	X
dealer								
American	-.018012	.004008	-4.49	0.000	-.025868	-.010157		18.976
Europe	-.00364	.000835	-4.36	0.000	-.005276	-.002003		3.461
Japan	.021652	.004631	4.68	0.000	.012576	.030728		7.5424
casevars								
_Isex_2	-.064827	.045389	-1.43	0.153	-.153787	.024133		.7322
income	.007061	.001564	4.52	0.000	.003996	.010125		42.097
size	-.081283	.01939	-4.19	0.000	-.119286	-.04328		2.6237

Pour estimer un modèle multinomial simple, on peut utiliser la ligne de commande suivante :

xi:asclogit choice, case(id) alternatives(car) casevars(i.sex income size)

Mais on pourrait penser utiliser la fonction suivante :

mlogit car sex income size (attention car et sex doivent être modifiées)

Au préalable, il faut modifier la base de données et ne garder que les choix = 1

keep if choice == 1

Remarque : à partir de cette base, il est possible de réaliser des statistiques

tab car

car	Freq.	Percent	Cum.
American	192	65.08	65.08
Japan	64	21.69	86.78
Europe	39	13.22	100.00
Total	295	100.00	

tab sex

sex	Freq.	Percent	Cum.
female	79	26.78	26.78
male	216	73.22	100.00
Total	295	100.00	

Comment passe d'une variable texte (string) à une variable numérique ?

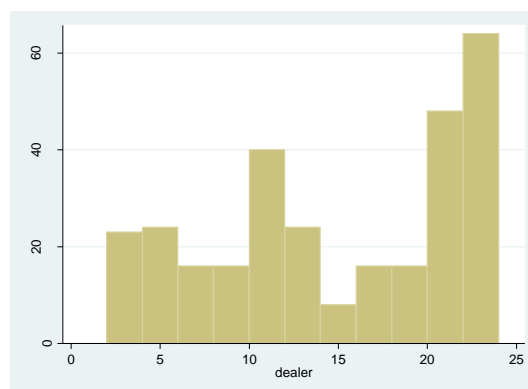
**encode car, generate(Car)
codebook Car**

```
type: numeric (long)
label: Car

range: [1,3]          units: 1
unique values: 3      missing .: 0/295

tabulation: Freq.   Numeric  Label
              192         1  American
               39         2  Europe
               64         3   Japan
```

histogram dealer, width(2) frequency



Estimation d'un modèle multinomiale simple à partir de la fonction mlogit :

mlogit Car _Isex_2 income size

```
Iteration 0: log likelihood = -259.1712
Iteration 1: log likelihood = -248.26305
Iteration 2: log likelihood = -247.95024
Iteration 3: log likelihood = -247.94956
Iteration 4: log likelihood = -247.94956
```

Multinomial logistic regression

Log likelihood = -247.94956

```
Number of obs    =      295
LR chi2(6)       =      22.44
Prob > chi2      =      0.0010
Pseudo R2       =      0.0433
```

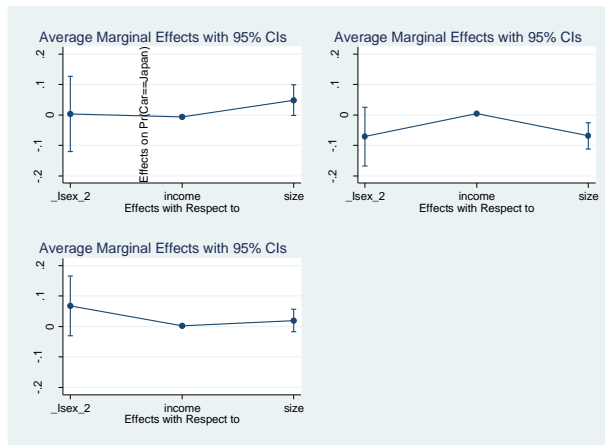
Qualité du modèle

Car	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
American	(base outcome)					
Europe						
_Isex_2	.5127444	.4552509	1.13	0.260	-.3795309	1.40502
income	.0265438	.0136208	1.95	0.051	-.0001524	.05324
size	.0758792	.1721839	0.44	0.659	-.2615952	.4133535
_cons	-3.334001	.8557467	-3.90	0.000	-5.011233	-1.656768
Japan						
_Isex_2	-.3546811	.3187085	-1.11	0.266	-.9793383	.2699762
income	.0309721	.0128444	2.41	0.016	.0057976	.0561466
size	-.4139465	.1453136	-2.85	0.004	-.6987559	-.1291371
_cons	-1.162682	.6862634	-1.69	0.090	-2.507734	.1823692

Pour calculer les odd-ratios :

mlogit Car _Isex_2 income size,rrr

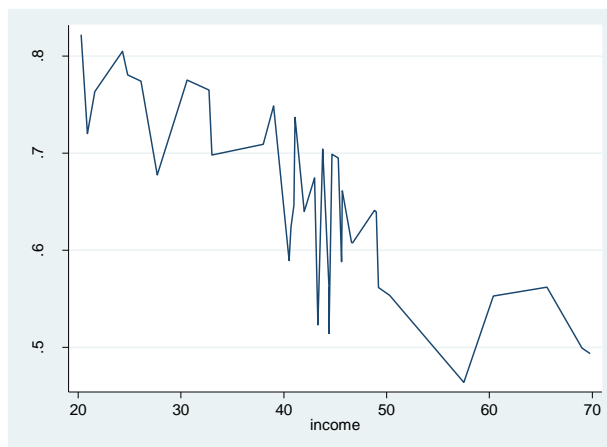
graph combine American Japan Europe, ycommon



71 Économétrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



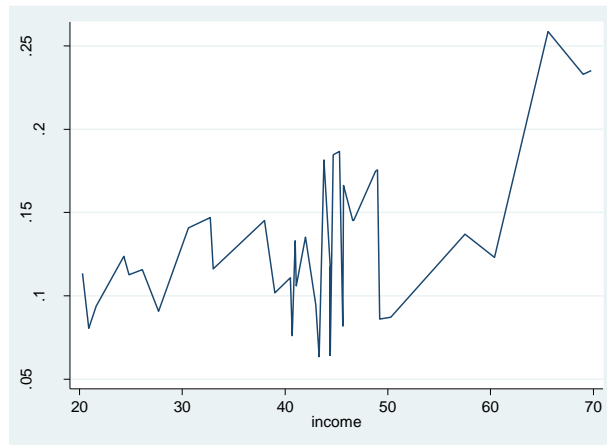
predict p1 p2 p3
sort income
twoway (line p1 income)



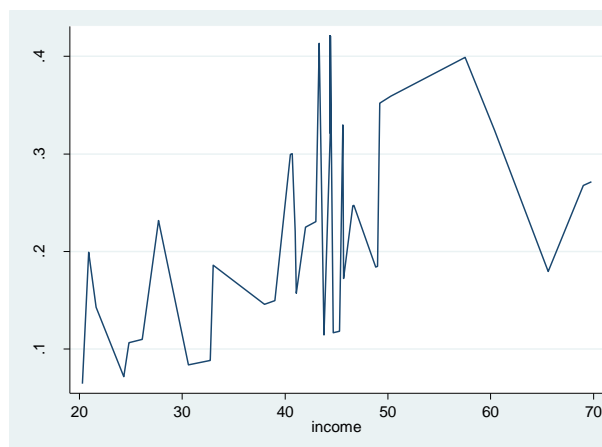
72 Économétrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



twoway (line p2 income)



twoway (line p3 income)



```
use "C:\Users\travers-
m\Desktop\Cours_2020_2021\Econometrie_variables_qualitatives_M1_EKAP_M2_COD
EME\Cours\Bases\mus15data.dta", clear
describe
```

```
obs:      1,182
vars:      16
size:      75,648
12 May 2008 20:46
```

variable name	storage type	display format	value label	variable label
mode	float	%9.0g	modetype	Fishing mode
price	float	%9.0g		price for chosen alternative
crate	float	%9.0g		catch rate for chosen alternative
dbeach	float	%9.0g		1 if beach mode chosen
dpier	float	%9.0g		1 if pier mode chosen
dprivate	float	%9.0g		1 if private boat mode chosen
dcharter	float	%9.0g		1 if charter boat mode chosen
pbeach	float	%9.0g		price for beach mode
ppier	float	%9.0g		price for pier mode
pprivate	float	%9.0g		price for private boat mode
pcharter	float	%9.0g		price for charter boat mode
qbeach	float	%9.0g		catch rate for beach mode
qpier	float	%9.0g		catch rate for pier mode
qprivate	float	%9.0g		catch rate for private boat mode
qcharter	float	%9.0g		catch rate for charter boat mode
income	float	%9.0g		monthly income in thousands \$

list in 1/3

1.	mode charter	price 182.93	crate .5391	dbeach 0	dpier 0	dprivate 0	dcharter 1	pbeach 157.93
	ppier 157.93	pprivate 157.93	pcharter 182.93	qbeach .0678	qpier .0503	qprivate .2601	qcharter .5391	income 7.083332
2.	mode charter	price 34.534	crate .4671	dbeach 0	dpier 0	dprivate 0	dcharter 1	pbeach 15.114
	ppier 15.114	pprivate 10.534	pcharter 34.534	qbeach .1049	qpier .0451	qprivate .1574	qcharter .4671	income 1.25
3.	mode private	price 24.334	crate .2413	dbeach 0	dpier 0	dprivate 1	dcharter 0	pbeach 161.874
	ppier 161.874	pprivate 24.334	pcharter 59.334	qbeach .5333	qpier .4522	qprivate .2413	qcharter 1.0266	income 3.75

list mode price crate pbeach ppier pprivate pcharter in 1/3, clean

	mode	price	crate	pbeach	ppier	pprivate	pcharter
1.	charter	182.93	.5391	157.93	157.93	157.93	182.93
2.	charter	34.534	.4671	15.114	15.114	10.534	34.534
3.	private	24.334	.2413	161.874	161.874	24.334	59.334

tabulate mode

Fishing mode	Freq.	Percent	Cum.
beach	134	11.34	11.34
pier	178	15.06	26.40
private	418	35.36	61.76
charter	452	38.24	100.00
Total	1,182	100.00	

Estimation d'un modèle multinomial simple en prenant l'alternative beach comme référence

mlogit mode income, baseoutcome(1)

Formatage de la base du format court en format long

generate id = _n

reshape long d p q, i(id) j(fishmode beach pier private charter) string

Data	wide	->	long
Number of obs.	1182	->	4728
Number of variables	17	->	9
j variable (4 values)		->	fishmode
xij variables:			
dbeach dpier ... dcharter		->	d
pbeach ppier ... pcharter		->	p
qbeach qpier ... qcharter		->	q

list in 1/8

	id	fishmode	mode	price	crate	d	p	q	income
1.	1	beach	charter	182.93	.5391	0	157.93	.0678	7.083332
2.	1	charter	charter	182.93	.5391	1	182.93	.5391	7.083332
3.	1	pier	charter	182.93	.5391	0	157.93	.0503	7.083332
4.	1	private	charter	182.93	.5391	0	157.93	.2601	7.083332
5.	2	beach	charter	34.534	.4671	0	15.114	.1049	1.25
6.	2	charter	charter	34.534	.4671	1	34.534	.4671	1.25
7.	2	pier	charter	34.534	.4671	0	15.114	.0451	1.25
8.	2	private	charter	34.534	.4671	0	10.534	.1574	1.25

asclogit d p q, case(id) alternatives(fishmode) casevars(income)

```

Alternative-specific conditional logit      Number of obs   =      4,728
Case variable: id                        Number of cases  =      1182

Alternative variable: fishmode            Alts per case: min =      4
                                           avg   =      4.0
                                           max   =      4

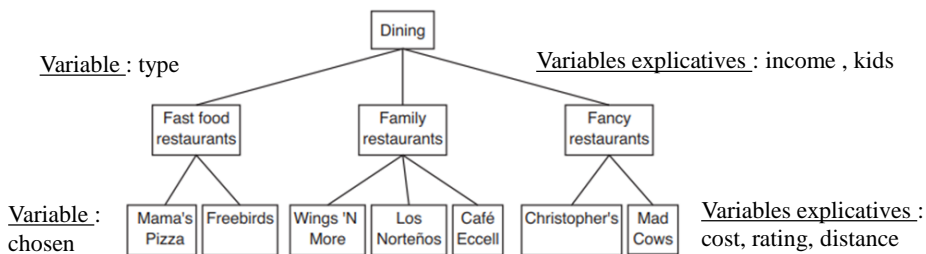
                                           Wald chi2(5)    =      252.98
Log likelihood = -1215.1376                Prob > chi2     =      0.0000

```

	d	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
fishmode						
	p	-.0251166	.0017317	-14.50	0.000	-.0285106 -.0217225
	q	.357782	.1097733	3.26	0.001	.1426302 .5729337
beach		(base alternative)				
charter						
	income	-.0332917	.0503409	-0.66	0.508	-.131958 .0653746
	_cons	1.694366	.2240506	7.56	0.000	1.255235 2.133497
pier						
	income	-.1275771	.0506395	-2.52	0.012	-.2268288 -.0283255
	_cons	.7779594	.2204939	3.53	0.000	.3457992 1.21012
private						
	income	.0894398	.0500671	1.79	0.074	-.0086898 .1875695
	_cons	.5272788	.2227927	2.37	0.018	.0906131 .9639444

B) Comment estimer un modèle « emboité » (nested model) sous Stata

Prenons l'exemple suivant :



use <http://www.stata-press.com/data/r13/restaurant>, clear
describe

Contains data from <http://www.stata-press.com/data/r13/restaurant.dta>

```

obs:      2,100
vars:      8                               10 Mar 2013 01:17
size:     67,200
  
```

variable name	storage type	display format	value label	variable label
family_id	float	%9.0g		family ID
restaurant	float	%12.0g	names	choices of restaurants
income	float	%9.0g		household income
cost	float	%9.0g		average meal cost per person
kids	float	%9.0g		number of kids in the household
rating	float	%9.0g		ratings in local restaurant guide
distance	float	%9.0g		distance between home and restaurant
chosen	float	%9.0g		0 no 1 yes

list family_id restaurant chosen kids rating distance cost income in 1/14, sepby(family_id)

	family~d	restaurant	chosen	kids	rating	distance	cost	income
1.	1	Freebirds	1	1	0	1.245553	5.444695	39
2.	1	MamasPizza	0	1	1	2.82493	6.19446	39
3.	1	CafeEccell	0	1	2	4.21293	8.182085	39
4.	1	LosNortenos	0	1	3	4.167634	9.861741	39
5.	1	WingsNmore	0	1	2	6.330531	9.667909	39
6.	1	Christophers	0	1	4	10.19829	25.95777	39
7.	1	MadCows	0	1	5	5.601388	28.99846	39
8.	2	Freebirds	0	3	0	4.162657	5.26874	58
9.	2	MamasPizza	0	3	1	2.865081	5.728618	58
10.	2	CafeEccell	0	3	2	5.337799	7.054855	58
11.	2	LosNortenos	1	3	3	4.282864	10.78514	58
12.	2	WingsNmore	0	3	2	8.133914	8.313948	58
13.	2	Christophers	0	3	4	8.664631	21.2801	58
14.	2	MadCows	0	3	5	9.119597	25.87567	58

nlogitgen type = restaurant(fast : Freebirds | MamasPizza , family : CafeEccell | LosNortenos | WingsNmore, fancy : Christophers | MadCows)

new variable type is generated with 3 groups

label list lb_type

lb_type:

- 1 fast
- 2 family
- 3 fancy

tree structure specified for the nested logit model

nlogittree restaurant type, choice(chosen)

type	N	restaurant	N	k
fast	600	Freebirds	300	12
		MamasPizza	300	15
family	900	CafeEccell	300	78
		LosNortenos	300	75
		WingsNmore	300	69
fancy	600	Christophers	300	27
		MadCows	300	24
		total	2100	300

k = number of times alternative is chosen

N = number of observations at each level

RUM-consistent nested logit regression	Number of obs	=	2,100
Case variable: family_id	Number of cases	=	300
Alternative variable: restaurant	Alts per case: min	=	7
	avg	=	7.0
	max	=	7
	Wald chi2(7)	=	46.71
Log likelihood = -485.47331	Prob > chi2	=	0.0000

chosen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
restaurant						
cost	-.1843847	.0933975	-1.97	0.048	-.3674404	-.0013289
rating	.463694	.3264935	1.42	0.156	-.1762215	1.10361
distance	-.3797474	.1003828	-3.78	0.000	-.5764941	-.1830007

fast						
income	-.0266038	.0117306	-2.27	0.023	-.0495952	-.0036123
kids	-.0872584	.1385026	-0.63	0.529	-.3587184	.1842016

family		
income	0	(base)
kids	0	(base)

fancy						
income	.0461827	.0090936	5.08	0.000	.0283595	.0640059
kids	-.3959413	.1220356	-3.24	0.001	-.6351267	-.1567559

type				
/fast_tau	1.712878	1.48685	-1.201295	4.627051
/family_tau	2.505113	.9646351	.614463	4.395763
/fancy_tau	4.099844	2.810123	-1.407896	9.607583

LR test for IIA (tau = 1): chi2(3) = 6.87 Prob > chi2 = 0.0762

→ Après vérification, il existe des variables explicatives corrélées entre elles

correlate cost rating distance income kids

	cost	rating	distance	income	kids
cost	1.0000				
rating	0.9070	1.0000			
distance	0.6445	0.6845	1.0000		
income	0.0007	0.0000	0.0049	1.0000	
kids	-0.0026	-0.0000	-0.0096	-0.0591	1.0000

→ Réestimation du modèle emboîté en ne conservant que la distance

**nlogit chosen distance || type: income kids, base(family) ||restaurant:, noconstant
case(family_id)**

```

RUM-consistent nested logit regression      Number of obs   =      2,100
Case variable: family_id                    Number of cases  =        300

Alternative variable: restaurant             Alts per case: min =         7
                                              avg =         7.0
                                              max =         7

Log likelihood = -488.15228                  Wald chi2(5)     =      44.73
                                              Prob > chi2      =      0.0000

```

chosen	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
restaurant						
distance	-.3759665	.0945169	-3.98	0.000	-.5612163	-.1907167
type equations						
fast						
income	-.0257008	.0117835	-2.18	0.029	-.0487961	-.0026055
kids	-.0741824	.1409697	-0.53	0.599	-.3504778	.2021131
family						
income	0	(base)				
kids	0	(base)				
fancy						
income	.0429303	.0088578	4.85	0.000	.0255692	.0602913
kids	-.3907581	.1193032	-3.28	0.001	-.6245881	-.156928
dissimilarity parameters						
type						
/fast_tau	2.537008	1.622585			-.6431995	5.717215
/family_tau	3.259947	1.018783			1.263169	5.256725
/fancy_tau	2.420526	1.586606			-.6891651	5.530217
LR test for IIA (tau = 1):						
			chi2(3) =	18.02	Prob > chi2 = 0.0004	

Si on reprend le cas de la base Transport

```
import excel "C:\Users\travers-  
m\Desktop\Cours_2020_2021\Econometrie_variables_qualitatives_M1_EKAP_M2_C  
ODEME\Cours\Transport.xls", sheet("Feuil1") firstrow clear  
describe
```

Contains data

```
obs:      840  
vars:      9  
size:    14,280
```

variable name	storage type	display format	value label	variable label
individual	int	%10.0g		individual
mode	str5	%9s		mode
choice	byte	%10.0g		choice
wait	byte	%10.0g		wait
vcost	int	%10.0g		vcost
travel	int	%10.0g		travel
gcost	int	%10.0g		gcost
income	byte	%10.0g		income
size	byte	%10.0g		size

list individual mode choice wait vcost travel gcost income in 1/12, sepby(individual)

	indivi~1	mode	choice	wait	vcost	travel	gcost	income
1.	1	air	0	69	59	100	70	35
2.	1	train	0	34	31	372	71	35
3.	1	bus	0	35	25	417	70	35
4.	1	car	1	0	10	180	30	35
5.	2	air	0	64	58	68	68	30
6.	2	train	0	44	31	354	84	30
7.	2	bus	0	53	25	399	85	30
8.	2	car	1	0	11	255	50	30
9.	3	air	0	69	115	125	129	40
10.	3	train	0	34	98	892	195	40
11.	3	bus	0	35	53	882	149	40
12.	3	car	1	0	23	720	101	40

```
nlogitgen type = mode(Aerien: air, Terrestre: bus | car | train)
```

new variable type is generated with 2 groups

label list lb_type

lb_type:

```
1 Aerien
2 Terrestre
```

```
nlogit choice wait vcost travel|| type: income size, base(Aerien) ||mode:, noconstant
case(individual)
```

note: branch 1 of level 1 is degenerate and the associated dissimilarity parameter ([Aerien_tau]_cons) is not defined; see help [nlogit](#) for details

tree structure specified for the nested logit model

type	N	mode	N	k
Aerien	210	air	210	58
Terrestre	630	bus	210	30
		car	210	59
		train	210	63
total			840	210

k = number of times alternative is chosen

N = number of observations at each level

91 Econométrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



```
Iteration 0: log likelihood = -250.7477
Iteration 1: log likelihood = -242.29619 (backed up)
Iteration 2: log likelihood = -241.15974 (backed up)
Iteration 3: log likelihood = -240.68196 (backed up)
Iteration 4: log likelihood = -240.67949 (backed up)
Iteration 5: log likelihood = -236.45105
Iteration 6: log likelihood = -236.23812
Iteration 7: log likelihood = -234.94851
Iteration 8: log likelihood = -234.75033
Iteration 9: log likelihood = -234.64708
Iteration 10: log likelihood = -234.64078
Iteration 11: log likelihood = -234.64001
Iteration 12: log likelihood = -234.63998
Iteration 13: log likelihood = -234.63998
```

RUM-consistent nested logit regression	Number of obs	=	840
Case variable: individual	Number of cases	=	210

Alternative variable: mode	Alts per case: min	=	4
	avg	=	4.0
	max	=	4

Log likelihood = -234.63998	Wald chi2(5)	=	46.16
	Prob > chi2	=	0.0000

92 Econométrie des variables qualitatives I – M.TRAVERS – Année 2020-2021



choice	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mode						
wait	-.0601601	.010929	-5.50	0.000	-.0815806	-.0387396
vcost	.026249	.0100951	2.60	0.009	.0064629	.0460351
travel	-.0027675	.0007437	-3.72	0.000	-.0042252	-.0013098
type equations						
Aerien						
income	0	(base)				
size	0	(base)				
Terrestre						
income	-.0392629	.0106161	-3.70	0.000	-.0600701	-.0184558
size	.4944531	.1995406	2.48	0.013	.1033607	.8855456
dissimilarity parameters						
type						
/Aerien_tau	1	970571			-1902283	1902285
/Terrestre_u	1.954363	.3835601			1.202599	2.706127
LR test for IIA (tau = 1):						
			chi2(2) =	8.91	Prob > chi2 = 0.0116	