CHAPITRE 4-2

LE MODÈLE LINÉAIRE GÉNÉRAL ET LA RÉGRESSION MULTIPLE APPLIQUÉS À L'ANALYSE DE VARIANCE

Plan

CHAPITRE 4-2
LE MODÈLE LINÉAIRE GÉNÉRAL ET LA RÉGRESSION MULTIPLE APPLIQUÉS À L'ANALYSE DE VARIANCE

Réf. : Wonnacott et Wonnacott (1992, p. 505-507) ; Iman et Conover (1989, chap. 16-17)

Nous avons vu que l'analyse de régression est une méthode statistique qui s'applique lorsqu'un modèle théorique propose une relation entre une variable dépendante continue et une ou plusieurs variables indépendantes continues ou discrètes. L'analyse de variance (*analysis of variance*, *ANOVA*), quant à elle, s'applique lorsque les variables indépendantes sont toutes discrètes. Ses caractéristiques marquantes sont donc

- un modèle de type stimulus-réaction
- un réaction mesurée au moyen d'une variable continue
- des stimuli mesurés au moyen de variables discrètes [1]

(si les stimuli sont mesurés au moyen de variables dont certaines sont discrètes et d'autres, continues, il s'agit d'analyse de *covariance* ; on doit alors utiliser le modèle linéaire général)

La question de recherche qui se pose généralement est : la réaction au stimulus est-elle significativement différente d'une catégorie à l'autre ?

Il existe des procédures spécifiques et des formats standard de présentation des résultats pour l'analyse de variance. Mais on peut réaliser de façon équivalente une analyse de variance au moyen de la régression linéaire. Cela comporte quelques avantages. D'abord, l'analyse de régression, contrairement à l'analyse de variance, n'impose pas de restrictions quant au plan d'échantillonnage (nombre d'observations par catégorie). Ensuite, on peut combiner l'analyse de la variance avec des variables indépendantes continues (on appelle parfois ce type de modèle un modèle d'analyse de *covariance*). Enfin, l'analyse de régression offre une plus grande flexibilité quant aux hypothèses que l'on peut soumettre à des tests statistiques.

4-2.1. Un exemple

Dans le cadre de la construction d'une matrice de comptabilité sociale pour le Québec, Robichaud *et al.* (1998) ont étudié l'épargne des ménages Québécois, à partir de données du

---

1 Selon le contexte et la discipline, on utilise différents termes utilisés pour désigner les variables indépendantes : facteurs, effets, catégories, variables qualitatives, variables de classement (*classification variables*), etc.

fichier de micro-données à grande diffusion de l'enquête de Statistique Canada sur les dépenses des familles en 1992. Le fichier contient 1900 observations pour le Québec.

Les six variables suivantes ont été tirées du fichier de micro-données à grande diffusion de l'enquête de Statistique Canada sur les dépenses des familles en 1992 :

1. Composition du ménage
   - Personnes seules
   - Couples [2] sans enfants
   - Couples avec enfants [3]
   - Familles monoparentales
   - Autres ménages [4]
2. Nombre d'enfants de moins de 16 ans
3. Âge de la personne de référence [5]
4. Revenu du ménage après impôt
5. Variation nette de l'actif et du passif
6. Sécurité [6]

La formulation du modèle linéaire dont les paramètres seront estimés repose sur le modèle conceptuel suivant. Le montant d'épargne d'un ménage (variable dépendante) augmente avec le revenu après impôt, mais dépend de l'âge du ménage (hypothèse du cycle de vie) et de la présence d'enfants (dépenses plus élevées) ; l'épargne peut aussi être influencée par le fait que la responsabilité du ménage soit assumée par une personne seule (ce qui implique généralement qu'il n'y a qu'un seul revenu, mais aussi, dans un grand nombre de cas, qu'il n'y a pas d'autre adulte à faire vivre) ; on voudra également vérifier si la catégorie hétéroclite des « Autres ménages » est différente.

---

[2] Mariés ou conjoints de fait.

[3] Par rapport à la composition du ménage, il s'agit d'enfants d'âge quelconque, jamais mariés et qui vivent avec leurs parents.

[4] Cette catégorie comprend les couples sans enfants vivant avec une personne apparentée qui n'est pas leur enfant, ainsi que les ménages dont au moins un membre n'est pas apparenté à la « personne de référence » (voir la note suivante). Se retrouvent notamment dans cette catégorie hétéroclite les ménages sans enfants avec chambreur et les groupes d'étudiants qui partagent un appartement.

[5] Dans l'enquête sur les dépenses des familles, la « personne de référence » est le membre du ménage qui est désigné par le répondant comme le principal soutien financier, habituellement, la personne dont le revenu est le plus élevé.

[6] Primes d'assurance-vie, etc.

Il faut signaler que le choix et la définition des variables indépendantes étaient dictés par l'objectif de la construction d'une matrice de comptabilité sociale et que le modèle qui est présenté ici ne doit pas être considéré comme un exemple de modélisation du comportement d'épargne des ménages. Il faut seulement y voir une illustration de la manière de traiter des variables indépendantes catégoriques dans la régression linéaire. Ajoutons que la présence du revenu parmi les variables indépendantes fait qu'il serait impossible d'appliquer à ce modèle une analyse de la variance classique (à moins de remplacer la variable de revenu par une variable catégorique) : ce modèle illustre donc aussi la plus grande polyvalence de l'analyse de régression.

Nous avons donc défini la variable dépendante

EPARGNE = Variation nette de l'actif et du passif + Sécurité

Les variables indépendantes sont

REVAPIMP = Revenu du ménage après impôt

Âge

Composition du ménage

Nous verrons dans un moment comment nous avons spécifié les variables d'âge et de composition du ménage. En combinant les variables catégoriques, on obtient une répartition des 1900 observations selon la composition du ménage et l'âge de la personne de référence. Cette répartition est donnée au tableau suivant, où l'on peut voir que les données ne se conforment pas à un plan d'échantillonnage « équilibré » (avec le même nombre d'observations dans chaque cellule), ni même à un plan d'échantillonnage bi-proportionnel. Dans ces conditions, il serait difficile de faire une analyse de variance classique ; mais l'utilisation de la régression multiple ne comporte pas de restriction semblable quant à la structure des données. Il faut cependant noter que certaines cellules ne comportent qu'un petit nombre d'observations, ce qui appelle une certaine prudence dans l'interprétation des résultats.

STRUCTURE DE L'ÉCHANTILLON (nombre d'observations)

Composition du ménage	**Âge de la personne de référence**				**Total**
	Moins de 35	**35-45**	**45-65**	**65 et plus**	
Personnes seules	101	77	132	142	452
Couples sans enf.	95	49	177	133	454
Couples avec enf.	163	267	239	26	695
Familles mono.	30	71	48	10	159
Autres sans enf.	24	17	41	21	103
Autres avec enf.	8	16	12	1	37
Total	421	497	649	333	1900

VARIABLES INDÉPENDANTES D'ÂGE

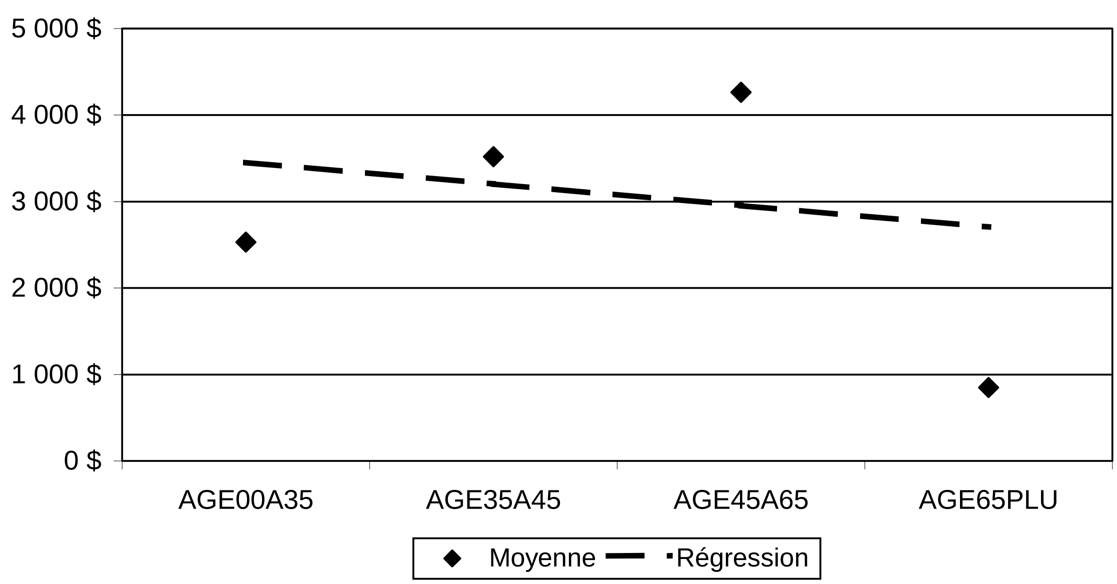
Nous avons d'abord défini la variable indépendante *GROUPAGE* :

Âge de la personne de référence	**Valeur de la variable *GROUPAGE***
moins de 35 ans	1
35 ans ou plus et moins de 45 ans	2
45 ans ou plus et moins de 65 ans	3
65 ans et plus	4

La variable *GROUPAGE* est une **variable ordinale d'ordre faible** (voir 1-1).

Mais cette variable ne peut pas figurer telle quelle dans la régression. Pourquoi ? Parce que cela imposerait artificiellement une relation linéaire entre le groupe d'âge et l'épargne, ce qui est contraire aux faits, comme l'illustre le graphique suivant.

Valeur moyenne de l'épargne par groupe d'âge et régression linéaire sur *GROUPAGE*



C'est pourquoi il faut remplacer GROUPAGE par des variables dichotomiques :

7. *AGE00A35* = 1 si *GROUPAGE* = 1 (âge < 35) ; = 0 autrement
8. *AGE35A45* = 1 si *GROUPAGE* = 2 (âge ≥ 35 et < 45) ; = 0 autrement
9. *AGE45A65* = 1 si *GROUPAGE* = 3 (âge ≥ 45 et < 65) ; = 0 autrement
10. *AGE65PLU* = 1 si *GROUPAGE* = 4 (âge ≥ 65) ; = 0 autrement

VARIABLES INDÉPENDANTES DE COMPOSITION DU MÉNAGE

La composition du ménage est une variable catégorique polytomique. Pour pouvoir distinguer entre les ménages « autres » avec, et sans enfants, nous utilisons l'information complémentaire donnée par le nombre d'enfants de moins de 16 ans. Nous obtenons ainsi 6 types de ménages :

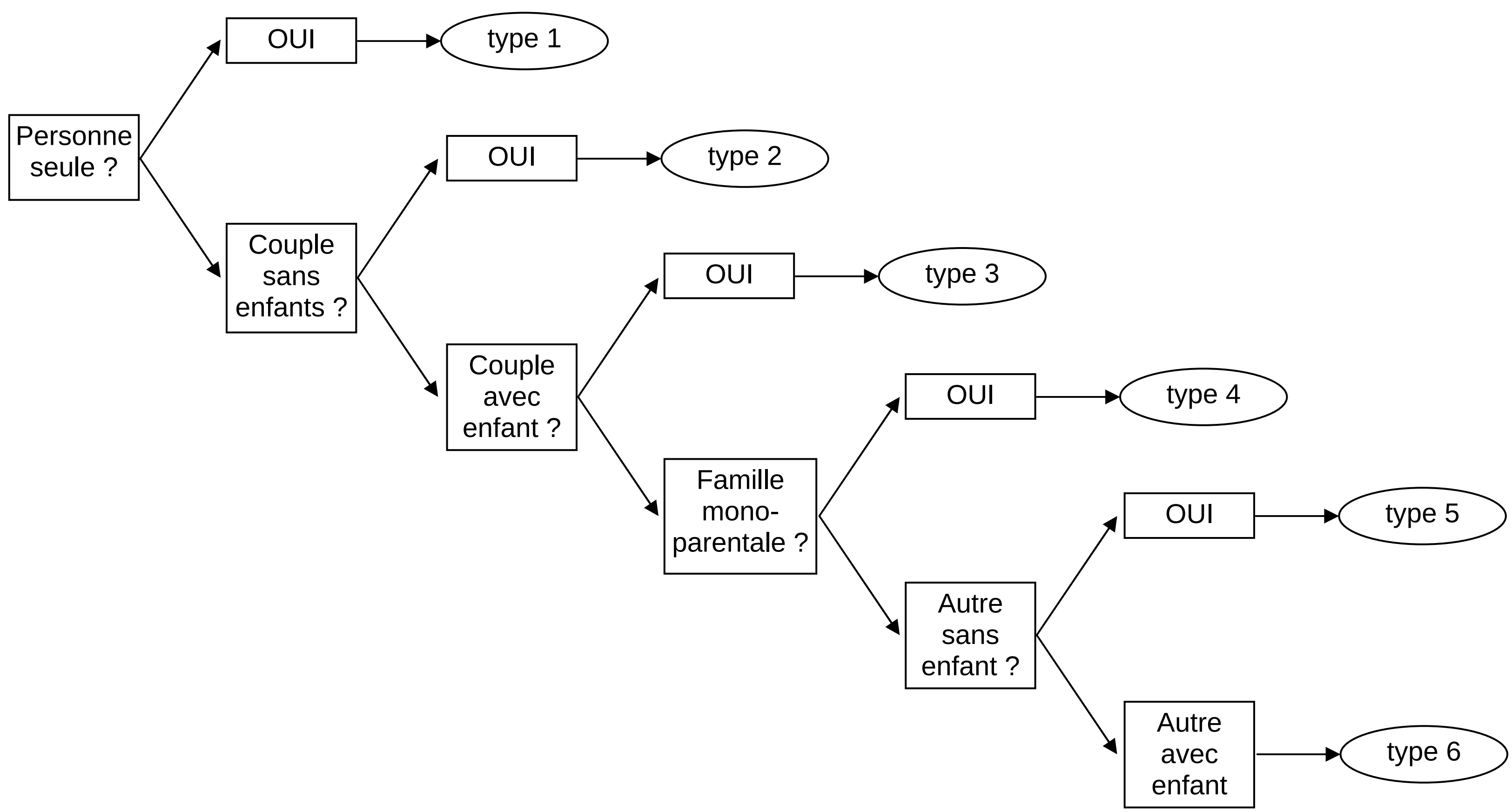
- Personnes seules
- Couples sans enf.
- Couples avec enf.
- Familles mono.
- Autres sans enf.
- Autres avec enf.

Mais la variable de composition du ménage, comme la variable d'âge, doit être remplacée par une série de variables dichotomiques, pour les mêmes raisons. Cela est encore plus nécessaire, du fait que la composition du ménage n'est même pas une variable ordinale (alors que *GROUPAGE* l'est).

Nous présentons ci-après deux approches de modélisation, l'une faisant appel à 5 variables dichotomiques et l'autre, à 3. Ces deux approches sont illustrées par les arbres de classification correspondants.

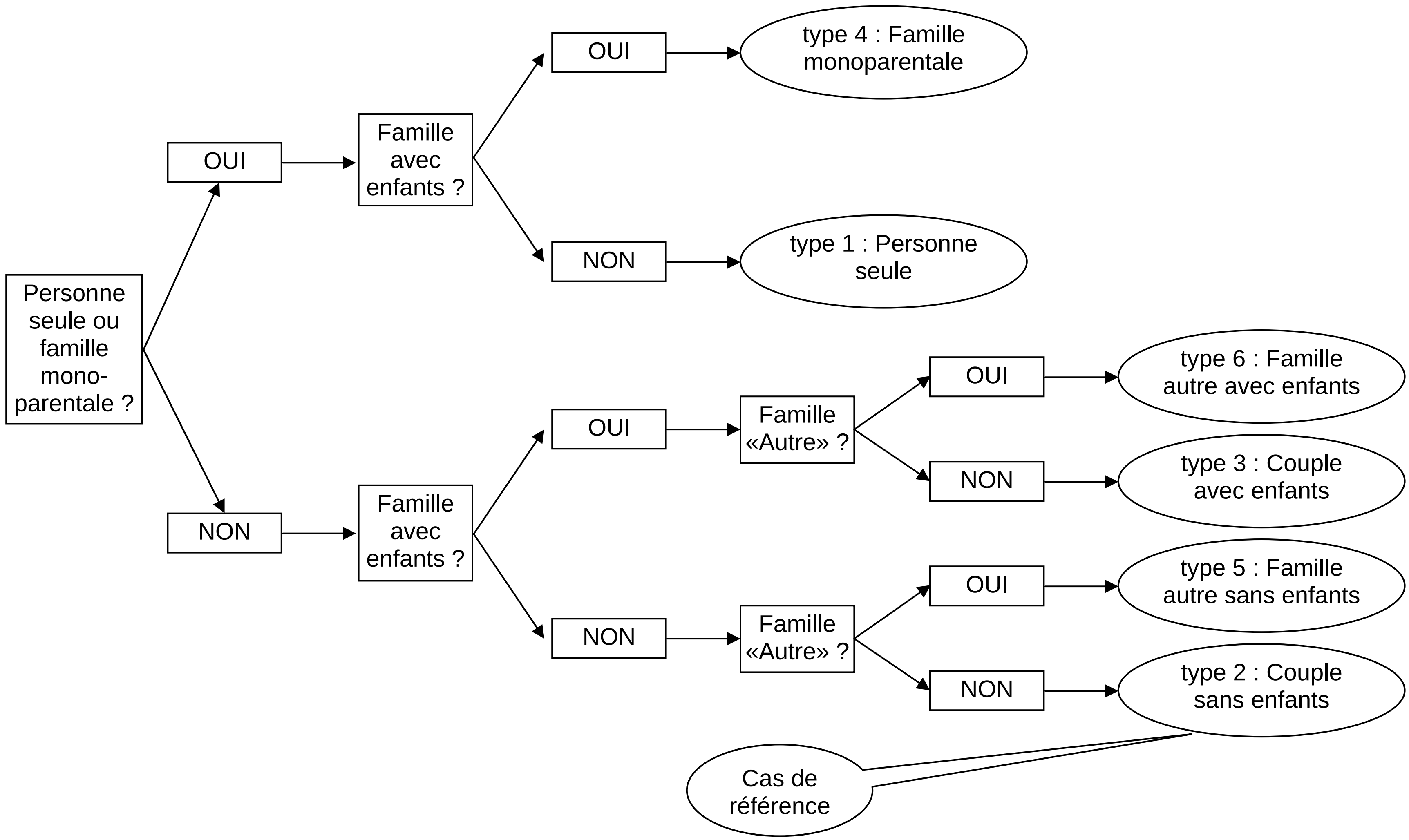
DEUX ARBRES DE CLASSIFICATION À L'AIDE DE VARIABLES DICHOTOMIQUES

I – Classification à 5 variables



DEUX ARBRES DE CLASSIFICATION À L'AIDE DE VARIABLES DICHOTOMIQUES

II – Classification à 3 variables



La classification à 5 variables consiste simplement à définir autant de variables dichotomiques qu'il y a de types de ménages, puis à laisser tomber l'une des variables, parce qu'elle est redondante (nous approfondirons ce point bientôt). Quant à la méthode à 3 variables, elle est constituée des variables dichotomiques suivantes :

1. *SEULMONO* = 1 s'il s'agit d'une personne seule ou d'une famille monoparentale ; *SEULMONO* = 0 autrement
2. *AUTRE* = 1 si le ménage appartient à la catégorie «Autres» ; *AUTRE* = 0 autrement
3. *ENFANTS* = 1 si le ménage compte au moins un enfant ; *ENFANTS* = 0 autrement

> Enfants = d'âge quelconque, jamais mariés et qui vivent avec leurs parents,
> SAUF pour les ménages «Autres» : enfants = de moins de 16 ans

C'est la classification à 3 variables que nous avons adoptée. En combinant ces trois variables dichotomiques, on obtient la classification suivante :

Composition du ménage	**Nombre d'enfants**	**Valeur de la variable *SEULMONO***	**Valeur de la variable *AUTRE***	**Valeur de la variable *ENFANTS***
Personnes seules	0	1	0	0
Couples sans enf.	0	0	0	0
Couples avec enf.	> 0	0	0	1
Familles mono.	> 0	1	0	1
Autres ménages	0	0	1	0
	> 0	0	1	1

On voit dans ce tableau que chaque type de ménage correspond à une combinaison particulière des variables dichotomiques.

Quelle est la différence entre les deux schèmes de classification ? Celui que nous avons adopté force, en quelque sorte, le modèle à un certaine cohérence, une cohérence qui ne reflète pas nécessairement la réalité. Par exemple, avec le triplet *SEULMONO*, *ENFANTS* et *AUTRE*, l'effet d'avoir des enfants doit être le même, quelles que soient les autres caractéristiques du ménage. Cela impose donc des restrictions au modèle. Nous verrons cependant que ces restrictions peuvent être relâchées par l'introduction de variables d'interaction (voir 4-2.4).

4-2.2. Élimination de la redondance parmi les variables indépendantes

Les variables dichotomiques d'âge ne doivent pas être incluses toutes les quatre parmi les variables indépendantes du modèle, parce que l'une de ces variables est redondante. En effet, si $AGE00A35 = 0$ *et* $AGE45A65 = 0$ *et* $AGE65PLU = 0$, alors, forcément, $AGE35A45 = 1$; en général, si, pour une observation donnée, trois des quatre variables ont la valeur zéro, la quatrième a nécessairement la valeur 1. Alors il faut écarter l'une des variables du modèle ; le cas qui correspond à la variable écartée devient le **cas de référence**. Nous choisissons la tranche d'âge de 35 à 45 comme cas de référence.

> Formellement, l'inclusion des quatre variables violerait la condition H4 du modèle classique de la régression linéaire, parce que leur somme est toujours égale à 1, c'est-à-dire à la variable constante du modèle :
>
> $AGE00A35 + AGE35A45 + AGE45A65 + AGE65PLU = 1 = CONSTANTE$

Il est à noter que nous avons déjà procédé implicitement à l'élimination de variables redondantes lorsque nous avons défini les variables dichotomiques *ENFANTS*, *SEULMONO* et *AUTRE*. Dans chaque cas, en effet, nous nous sommes abstenus de définir deux variables, une par catégorie. Par exemple, nous aurions pu définir

- $AVECENFANTS = 1$ si le ménage compte au moins un enfant ; $AVECENFANTS = 0$ autrement
- $SANSENFANTS = 0$ si le ménage compte au moins un enfant ; $SANSENFANTS = 1$ autrement

Nous ne l'avons pas fait, parce que l'une de ces deux variables aurait été redondante.

Nous avons montré en 4-2.1 comment la variable *Composition du ménage* a été remplacée par le triplet *SEULMONO*, *ENFANTS* et *AUTRE*. Si on avait voulu adopter l'autre schème de classification, il aurait fallu éliminer la redondance. C'est pourquoi le schème de classification comporte 5 variables dichotomiques pour 6 types de ménages.

4-2.3 Spécification d'un modèle sans interaction

Nous pouvons maintenant énoncer une première spécification du modèle :

$$EPARGNE = \beta_1 + \beta_2\ REVAPIMP + \beta_3\ SEULMONO + \beta_4\ AUTRE$$
$$+ \beta_5\ ENFANTS + \beta_6\ AGE00A35 + \beta_7\ AGE45A65 + \beta_8\ AGE65PLU$$

où on remarque l'absence de la variable *AGE35A45*, qui serait redondante.

Voyons maintenant ce que signifie ce modèle pour chacun des 24 cas de figure que distinguent nos données. Les 24 cas sont présentés dans le tableau intitulé « Interprétation du modèle sans interaction ».

Interprétation du modèle sans variables d'interaction

Composition du ménage	Nombre d'enfants	Groupe d'âge de la personne de référence	*SEULMONO*	*AUTRE*	*ENFANTS*	*AGE00A35*	*AGE45A65*	*AGE65PLU*	*EPARGNE* prédite par le modèle
Personnes seules	0	< 35	1	0	0	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + 0 + \beta_6 + 0 + 0$
		≥ 35 et < 45	1	0	0	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + 0 + 0 + 0 + 0$
		≥ 45 et < 65	1	0	0	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + 0 + 0 + \beta_7 + 0$
		≥ 65	1	0	0	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + 0 + 0 + 0 + \beta_8$
Couples sans enf.	0	< 35	0	0	0	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + 0 + \beta_6 + 0 + 0$
		≥ 35 et < 45	0	0	0	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + 0 + 0 + 0 + 0$
		≥ 45 et < 65	0	0	0	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + 0 + 0 + \beta_7 + 0$
		≥ 65	0	0	0	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + 0 + 0 + 0 + \beta_8$
Couples avec enf.	> 0	< 35	0	0	1	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + \beta_5 + \beta_6 + 0 + 0$
		≥ 35 et < 45	0	0	1	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + \beta_5 + 0 + 0 + 0$
		≥ 45 et < 65	0	0	1	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + \beta_5 + 0 + \beta_7 + 0$
		≥ 65	0	0	1	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + 0 + \beta_5 + 0 + 0 + \beta_8$
Familles mono.	> 0	< 35	1	0	1	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + \beta_5 + \beta_6 + 0 + 0$
		≥ 35 et < 45	1	0	1	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + \beta_5 + 0 + 0 + 0$
		≥ 45 et < 65	1	0	1	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + \beta_5 + 0 + \beta_7 + 0$
		≥ 65	1	0	1	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ \beta_3 + 0 + \beta_5 + 0 + 0 + \beta_8$
Autres ménages	0	< 35	0	1	0	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + 0 + \beta_6 + 0 + 0$
		≥ 35 et < 45	0	1	0	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + 0 + 0 + 0 + 0$
		≥ 45 et < 65	0	1	0	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + 0 + 0 + \beta_7 + 0$
		≥ 65	0	1	0	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + 0 + 0 + 0 + \beta_8$
	> 0	< 35	0	1	1	1	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + \beta_5 + \beta_6 + 0 + 0$
		≥ 35 et < 45	0	1	1	0	0	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + \beta_5 + 0 + 0 + 0$
		≥ 45 et < 65	0	1	1	0	1	0	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + \beta_5 + 0 + \beta_7 + 0$
		≥ 65	0	1	1	0	0	1	$\beta_1 + \beta_2$ *REVAPIMP* $+ 0 + \beta_4 + \beta_5 + 0 + 0 + \beta_8$

On peut voir dans ce tableau qu'à chacun des 24 cas possibles correspond une combinaison unique de valeurs des variables dichotomiques : il ne manque pas de variables, puisque tous les cas sont représentés distinctement. On remarque aussi que le cas de référence, celui pour lequel toutes les variables dichotomiques sont nulles, est celui d'un couple sans enfants dont la personne de référence a entre 35 et 45 ans. Il s'ensuit que les coefficients des variables dichotomiques représentent des différences par rapport à ce cas de référence ; par exemple, le modèle prédit qu'entre le ménage de référence et une famille monoparentale dont la personne responsable a moins de 35 ans, à revenu égal, il y aura un écart égal à $\beta_3 + \beta_5 + \beta_6$. On s'attendrait à ce que chacun de ces trois coefficients soit négatif, mais c'est l'estimation du modèle qui nous éclairera là-dessus.

Le tableau permet aussi de mieux voir pourquoi il aurait été absurde d'inclure dans le modèle la variable polytomique *GROUPAGE*.

Les résultats de l'estimation sont rapportés dans le tableau suivant.

Variable	**Description**	**Symbole**	**Coefficient estimé**	**t de Student**	**Probabilité critique**
CONSTANTE		β_1	-7727	-11,062	0,0001
REVAPIMP	Revenu du ménage après impôt	β_2	0,340	28,468	0,0001
ENFANTS	Présence d'enfants	β_5	-2260	-4,937	0,0001
SEULMONO	Pers.seule ou monop.	β_3	1903	3,834	0,0001
AUTRE	Ménage « Autre »	β_4	-2578	-3,309	0,0010
AGE00A35	Âge 00-35	β_6	258	0,444	0,6574
AGE45A65	Âge 45-65	β_7	419	0,796	0,4263
AGE65PLU	Âge 65+	β_8	875	1,322	0,1862

$n = 1900$

$R^2 = 0{,}33$

On constate notamment que les coefficients des variables relatives à l'âge ne sont pas significativement différents de zéro. Est-ce à dire que l'âge n'a pas d'effet sur le comportement d'épargne ?

4-2.4 Introduction des effets d'interaction

Le modèle présenté dans le tableau précédent ne tient pas compte de la possibilité d'effets d'interaction. Il y a un grand nombre d'interactions possibles et il est rare qu'un modèle les contienne toutes.

Par exemple, le modèle prédit que l'effet sur l'épargne de la présence d'enfants est égal à β_5, quels que soient l'âge de la personne de référence et la composition du ménage : est-ce bien comme cela ? En d'autres mots, n'y a-t-il pas *interaction* entre la variable *ENFANTS* et les variables *SEULMONO*, *AUTRE*, *AGE00A35*, *AGE45A65*, et *AGE65PLU* ? Il est à noter que chacun des effets d'interaction évoqués dans la phrase précédente est *symétrique* (comme le dénote le choix du mot « interaction ») : par exemple, au lieu de demander si l'effet de la présence d'enfants (*ENFANTS*) est modifié par l'appartenance au groupe des moins de 35 ans (*AGE00A35*), on peut se demander de façon équivalente si l'effet de l'appartenance au groupe des moins de 35 ans est modifié par la présence d'enfants.

Pour inclure la possibilité d'effets d'interaction dans le modèle, il faut ajouter des variables. Ainsi, on définit

4. *MONOMONO* = 1 si *ENFANTS* = 1 **et** *SEULMONO* = 1 ;
   MONOMONO = 0 autrement

Mathématiquement, et de façon plus concise, on définit [7]

11. *MONOMONO* = *ENFANTS* × *SEULMONO*

Dans la même veine, on a

12. *AUTRENFA* = *ENFANTS* × *AUTRE*
13. *ENFA0035* = *ENFANTS* × *AGE00A35*
14. *ENFA4565* = *ENFANTS* × *AGE45A65*
15. *ENFA65PL* = *ENFANTS* × *AGE65PLU*
16. *AUTA0035* = *AUTRE* × *AGE00A35*
17. *AUTA4565* = *AUTRE* × *AGE45A65*
18. *AUTA65PL* = *AUTRE* × *AGE65PLU*
19. *SOLA0035* = *SEULMONO* × *AGE00A35*

---

[7] Les variables dichotomiques sont des variable logiques, ou *booléennes*. En algèbre booléenne, la conjonction « et » se représente par la multiplication.

20. *SOLA4565* = *SEULMONO* × *AGE45A65*

21. *SOLA65PL* = *SEULMONO* × *AGE65PLU*

Dans la liste qui prédède, on peut voir que l'on n'a pas inclus toutes les variables possibles (par exemple, il n'y a aucune variable d'interaction avec *AGE35A45*). Car il en est des variables d'interaction comme des autres groupes de variables catégoriques : si l'on inclut dans le modèle toutes les variables possibles, il y a redondance.

Les coefficients des variables d'interaction s'interprètent comme des différences. Par exemple, nous avons vu dans le tableau de la section 4-2.3 que β_5, le coefficient de la variable *ENFANTS*, représente la différence, quant au montant épargné, entre deux ménages identiques à tous égards sauf pour la présence d'enfants ; de même, β_7, le coefficient de la variable *AGE45A65*, représente la différence, quant au montant épargné, entre deux ménages identiques à tous égards sauf pour l'âge, l'un appartenant au groupe d'âge de référence (35-45 ans) et l'autre, au groupe des 45-65 ans. Sans variable d'interaction, ces différences sont additives : le modèle de la section 4-2.3 prédit qu'entre un ménage sans enfants du groupe 35-45 ans et un ménage avec enfants du groupe 45-65 ans, la différence sera égale à $\beta_5 + \beta_7$. Si l'on ajoute au modèle la variable d'interaction *ENFA4565*, cette différence sera égale à $\beta_5 + \beta_7$, *plus le coefficient de la variable d'interaction ENFA4565* [8].

Il peut aussi y avoir interaction entre une variable catégorique et une variable continue. Ainsi, le modèle prédit que, quelles que soient les caractéristiques du ménage, un accroissement d'un dollar du revenu après impôt entraînera un accroissement de l'épargne de β_2 dollar. Se pourrait-il que cet effet soit différent pour les ménages avec enfants ? Pour examiner cette question dans le cadre du modèle, il faut y inclure des variables d'interaction. Nous en considérons trois. Les variables supplémentaires sont :

22. *REVENFAN* = *REVAPIMP* × *ENFANTS*

23. *REVSELMO* = *REVAPIMP* × *SEULMONO*

24. *REVAUTRE* = *REVAPIMP* × *AUTRE*

Les coefficients de ces variables s'interprètent aussi comme des différences. Par exemple, si l'on compare deux ménages identiques sauf pour la présence d'enfants, le coefficient de

[8] On peut faire une analogie avec la pharmacologie : l'effet d'une combinaison de deux médicaments peut se révéler fort différent des effets de chacun des médicaments pris séparément. Les médicaments pris ensemble peuvent se renforcer mutuelement ou au contraire s'annuler l'un l'autre.

REVENFAN représente la différence entre les deux quant à leur la propension marginale à épargner.

Après inclusion des variables d'interaction, le modèle complété devient donc

$$
\begin{aligned}
EPARGNE = {} & \beta_1 + \beta_2\ REVAPIMP + \beta_3\ SEULMONO + \beta_4\ AUTRE + \beta_5\ ENFANTS \\
& + \beta_6\ AGE00A35 + \beta_7\ AGE45A65 + \beta_8\ AGE65PLU \\
& + \gamma_1\ MONOMONO + \gamma_2\ AUTRENFA \\
& + \gamma_3\ ENFA0035 + \gamma_4\ ENFA4565 + \gamma_5\ ENFA65PL \\
& + \gamma_6\ AUTA0035 + \gamma_7\ AUTA4565 + \gamma_8\ AUTA65PL \\
& + \gamma_9\ SOLA0035 + \gamma_{10}\ SOLA4565 + \gamma_{11}\ SOLA65PL \\
& + \alpha_1\ REVENFAN + \alpha_2\ REVSELMO + \alpha_3\ REVAUTRE
\end{aligned}
$$

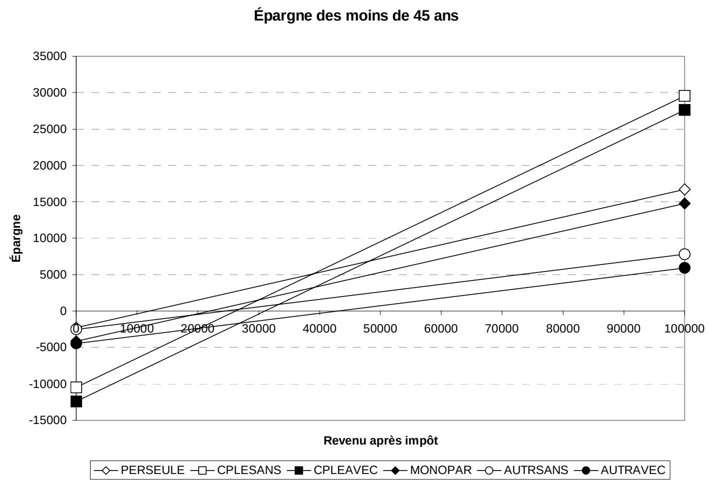
4-2.5 Estimation et interprétation du modèle

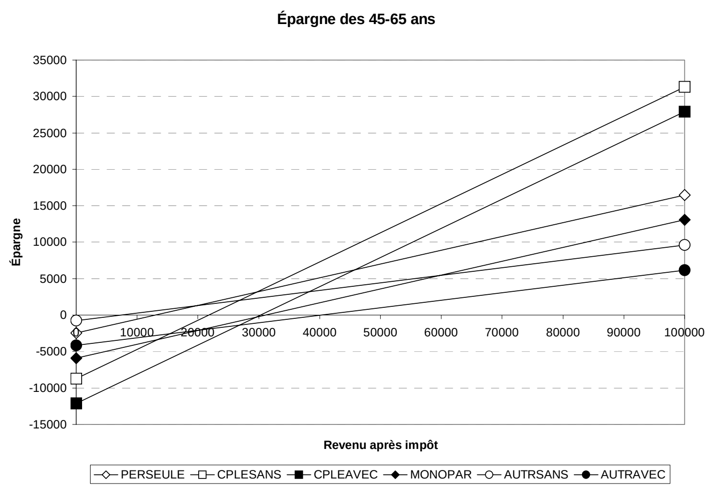
Après exécution de la procédure *backward* pour éliminer les variables dont les coefficients ne sont pas significatifs, on obtient les résultats qui sont présentés dans le tableau suivant.

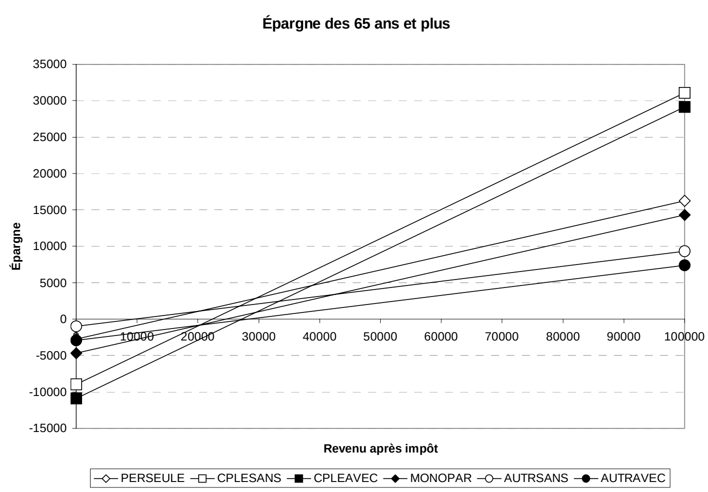
Variable	**Description**	**Symbole**	**Coefficient estimé**	**Erreur type**	**Probabilité critique**
CONSTANTE		β_1	-10487	729	0,0001
REVAPIMP	Revenu du ménage après impôt	β_2	0,400	0,013	0,0001
ENFANTS	Présence d'enfants	β_5	-1927	561	0,0006
SEULMONO	Pers.seule ou monoparentale	β_3	8233	1000	0,0001
AUTRE	Ménage « Autre »	β_4	7969	1771	0,0001
AGE45A65	Âge 45-65	β_7	1767	779	0,0234
AGE65PLU	Âge 65+	β_8	1513	770	0,0497
Variables d'interaction					
ENFA4565	*ENFANTS* × *AGE45A65*	γ_4	-1506	897	0,0932
SOLA4565	*SEULMONO* × *AGE45A65*	γ_{10}	-1983	1008	0,0494
SOLA65PL	*SEULMONO* × *AGE65PLU*	γ_{11}	-1996	1138	0,0796
REVSELMO	*REVAPIMP* × *SEULMONO*	α_2	-0,211	0,030	0,0001
REVAUTRE	*REVAPIMP* × *AUTRE*	α_3	-0,297	0,046	0,0001

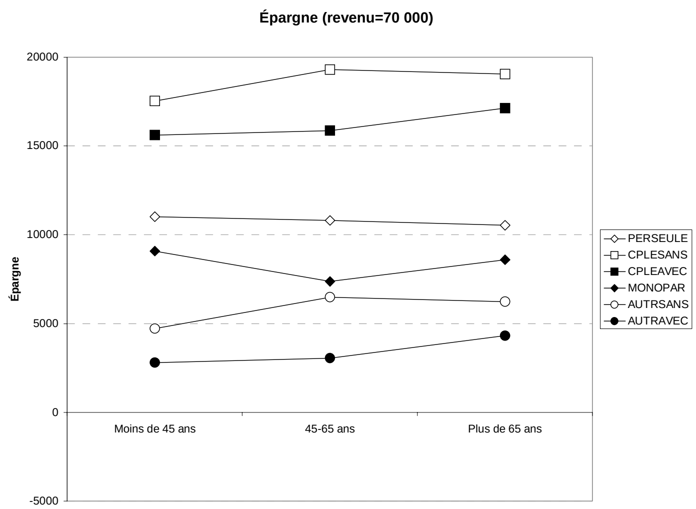
Le nombre d'observations est de 1900 et le coefficient de détermination multiple R^2 est de 0,36.

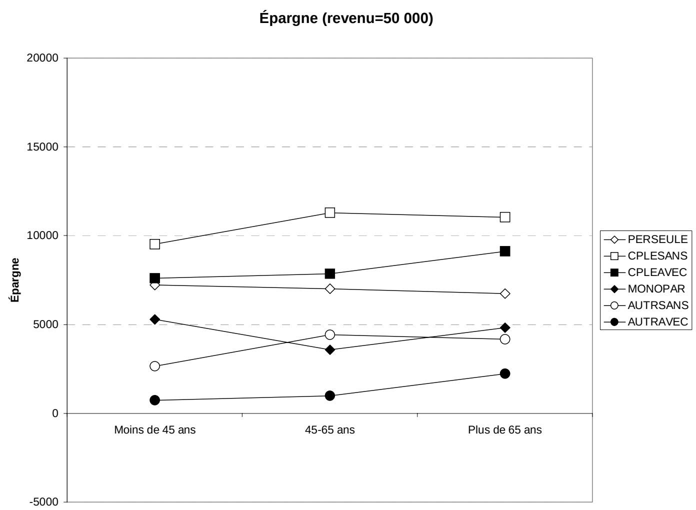
Il n'est pas aisé de voir clair dans tous ces coefficients : qu'est-ce que cela signifie au juste ? Les figures qui suivent illustrent les prédictions du modèle.

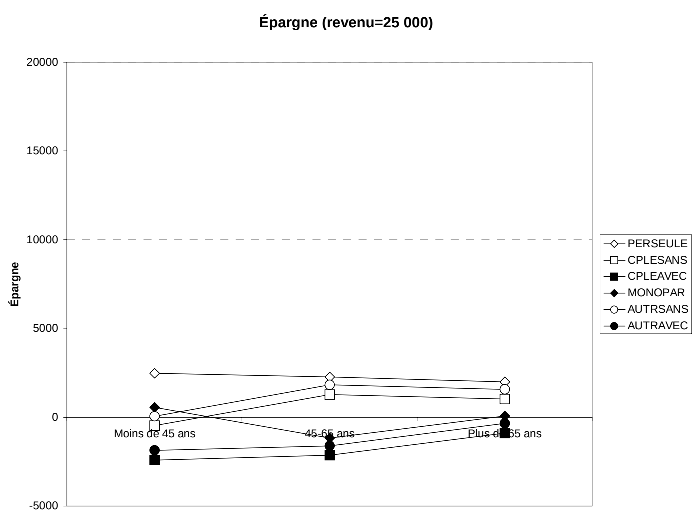

Épargne des moins de 45 ans
35000
30000
25000
20000
15000
10000
5000
0
-5000
-10000
-15000
Épargne
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Revenu après impôt
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC


Épargne des 45-65 ans
35000
30000
25000
20000
15000
10000
5000
0
-5000
-10000
-15000
Épargne
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Revenu après impôt
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC


Épargne des 65 ans et plus
35000
30000
25000
20000
15000
10000
5000
0
-5000
-10000
-15000
Épargne
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
Revenu après impôt
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC


Épargne (revenu=70 000)
20000
15000
10000
5000
0
-5000
Épargne
Moins de 45 ans
45-65 ans
Plus de 65 ans
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC


Épargne (revenu=50 000)
20000
15000
10000
5000
0
-5000
Épargne
Moins de 45 ans
45-65 ans
Plus de 65 ans
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC


Épargne (revenu=25 000)
20000
15000
10000
5000
0
-5000
Épargne
Moins de 45 ans
45-65 ans
Plus de 65 ans
PERSEULE
CPLESANS
CPLEAVEC
MONOPAR
AUTRSANS
AUTRAVEC