L'économétrie et l'étude des comportements : présentation et mise en œuvre de modèles de régression qualitatifs, les modèles à résidus logistiques ou normaux (LOGIT, PROBIT)

Dominique Leblanc, Stéphane Lollivier, Maryse Marpsat, Daniel Verger

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES

Série des documents de travail

« Méthodologie Statistique »

N° 0001

L'ECONOMETRIE ET L'ETUDE DES COMPORTEMENTS

Présentation et mise en oeuvre de modèles de régression qualitatifs

Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)

Ce texte est une version actualisée et complétée du document (N° 9606)

Cette note est le fruit d'un travail collectif auquel ont participé G. GRIMLER, D. LE BLANC, S. LOLLIVIER, M. MARPSAT, H. ROUSSE, A. TROGNON, D. VERGER. Ce travail a bénéficié des remarques de A. JACQUOT et L. TOULEMON.

Les rédacteurs en sont D. LE BLANC, S. LOLLIVIER, M. MARPSAT et D. VERGER. Leur adresser vos remarques, suggestions, corrections, critiques, afin d'améliorer cette version.

L' ÉCONOMÉTRIE ET L'ÉTUDE DES COMPORTEMENTS
Présentation et mise en œuvre de modèles de régression qualitatifs
Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)

D. LEBLANC

INSEE - Direction des Statistiques Démographiques et Sociales

Département des prix à la consommation, des ressources et des conditions de vie des ménages

S. LOLLIVIER

GENES - ENSAE

M. MARPSAT

INED

D. VERGER

INSEE - Direction des Statistiques Démographiques et Sociales

Unité de Méthodologie Statistique

RÉSUMÉ

Ce document présente certaines méthodes économétriques de régression sur variables qualitatives. Il est essentiellement consacré à l'étude des modèles à variable dépendante dichotomique (logit ou probit selon les hypothèses sur les résidus). Le cas des modèles polytomiques ordonnés est aussi abordé, mais plus sommairement. Dans un premier temps, nous précisons les particularités des modèles à réponses qualitatives par rapport aux modèles traditionnels d'analyse de variance. Nous montrons ensuite que, bien que relevant d'une même logique, la modélisation est plus complexe puisqu'elle fait intervenir une variable latente. C'est sur cette dernière qu'est postulé le modèle linéaire habituel. Nous expliquons enfin les méthodes de résolution de ces modèles, ainsi que les tests pouvant être mis en œuvre.

Dans un second temps, nous expliquons comment ces modèles peuvent être estimés par le logiciel SAS. Sont fournis en particulier quelques suggestions sur le choix des variables explicatives, certaines mises en garde sur les difficultés liées à l'interprétation, et des conseils sur la lecture des résultats. Le texte se termine par l'exposé de quelques problèmes économétriques fréquemment ignorés.

MOTS CLÉS : Variables qualitatives ; modèles Logit et Probit ; économétrie sur variables latentes.

L'ECONOMETRIE ET L'ETUDE DES COMPORTEMENTS

Présentation et mise en oeuvre de modèles de régression qualitatifs

Les modèles à résidus logistiques ou normaux (dits LOGIT, PROBIT)

SOMMAIRE

pages

I Les variables qualitatives 9

1. Les variables dichotomiques
2. Les variables polytomiques
   a . Les variables polytomiques ordonnées
   b. Les variables non ordonnées

II Pourquoi des modèles particuliers ? 10

III Niveau d'utilité, variables latentes 11

IV Le (s) modèle (s) théorique (s) 12

V Les modèles PROBIT et LOGIT 13

1. Le modèle PROBIT
2. Le modèle LOGIT
3. Comparaison des deux modèles

VI L'estimation : formules, précisions techniques 15

1. Le principe de la méthode
2. L'algorithme utilisé
3. Quelques propriétés asymptotiques de l'estimateur du maximum de vraisemblance

pages

VII Les tests *19*

1. Test de la nullité d'un coefficient
2. Test d'une liaison de la forme : $\sum_{k=1}^{I} \lambda_k \beta_k = C$
3. Test de la nullité d'une ensemble de coefficients
   - test de Wald
   - test du rapport de vraisemblance
4. Cas plus général : test d'une hypothèse linéaire de la forme $Q'\beta = C$
5. Test de la validité générale du modèle

VIII Mise en oeuvre de la procédure LOGISTIC de la version 6 de SAS *29*

1. Quelques remarques et mises en garde préalables
2. Quelques rappels de syntaxe
3. Quelques précisions sur les procédures de sélection pas à pas des variables
4. Un exemple de sortie interprétée
5. Le fichier en sortie
6. Modèle LOGIT, modèle PROBIT

IX Mise en oeuvre du modèle LOGIT *47*

1. La spécification du modèle
   a. retenir ou non une dimension explicative
   b. représentation d'une dimension explicative retenue
      - quelles variables pour une dimension ?
      - la situation de référence : à quoi sert-elle ? Comment la choisir ?
   c. introduction simultanée de plusieurs dimensions explicatives : problèmes spécifiques à éviter
      - les problèmes de colinéarité
      - les défauts d'additivité
   d. les variables omises
   e. pondérer ou ne pas pondérer : that's the question !
   f. l'endogénéïté
2. La lecture des résultats
   a. significativité des coefficients
   b. l'interprétation des coefficients en termes de probabilité
   c. significativité globale d'une dimension explicative
   d. peut-on classer les diverses dimensions explicatives par ordre d'importance (puissance explicative) ?
   e. les coefficients égaux à $\pm\infty$
   f. derniers problèmes
3. La publication des résultats

pages

X Quelques problèmes économétriques souvent ignorés *83*

1. L'hétéroscédasticité
2. L'asymétrie de la distribution des perturbations
3. Test de mauvaise spécification

XI Extension au cas d'une variable dépendante polytomique ordonnée *86*

Conclusion *88*

Bibliographie *89*

Préambule

Une analyse des comportements court le risque de rester incomplète si on se limite à l'observation de tableaux croisés ventilant une pratique selon un ou plusieurs critères. En effet, divers effets de structure peuvent conduire à des interprétations erronées ; il est alors nécessaire d'isoler les effets propres de telle ou telle variable.

Pour ce faire, les tabulations croisées son en général insuffisantes : même pour des enquêtes dont l'échantillon est grand, on se heurte très vite aux problèmes que pose le grand nombre de cases qui ne regroupent qu'un effectif très faible de ménages.

Pour aller plus loin, et tenter d'isoler l'effet spécifique d'un facteur « toutes choses égales par ailleurs », il faut faire des hypothèses et postuler des régularités statistiques.

Quand le phénomène étudié est continu (exemple : le revenu ou son logarithme, la consommation ou son logarithme), la méthode appropriée est **l'analyse de la variance**. Cette méthode est une extension naturelle du modèle de régression par les moindres carrés ordinaires, ou MCO.

Toutefois, dans une étude sur le comportement des ménages ou des individus, les pratiques étudiées sont le plus souvent de nature discrète, qualitative. Le recours à une analyse économétrique d'un type particulier est alors nécessaire pour isoler les effets propres (on parlera aussi de « séparation des effets », d'« effet d'une variable toutes choses égales par ailleurs », ou d'« effet d'une variable conditionnellement aux variables introduites dans le modèle »).

La procédure SAS décrite dans ce document correspond à celle de la version 6.12 sous Windows. Elle est appelée à évoluer dans les versions ultérieures.

I Les variables qualitatives

Différents types de variables qualitatives se rencontrent fréquemment dans nos enquêtes auprès des ménages :

1. Les variables dichotomiques

Ce sont des variables qui prennent deux valeurs, on dira aussi qui ont deux modalités, souvent notées 0 et 1.

Exemples :

- la possession d'un bien durable : 1 si le ménage possède le bien
  0 s'il ne le possède pas
- la pratique d'une activité : 1 si l'individu pratique l'activité
  0 s'il ne la pratique pas

2. Les variables polytomiques

Ce sont des variables qui prennent plus de deux valeurs ou modalités.
On distingue deux sortes de variables polytomiques, qui seront traitées différemment dans les modèles :

a. les variables polytomiques ordonnées

Les différentes modalités sont ordonnées dans un ordre « naturel », qu'elles proviennent d'un classement selon une intensité (fréquence d'activité, nombre de biens durables,..), ou qu'elles résultent du découpage en tranches d'une variable numérique (revenu,...).

Exemple : faire du sport
1. Tous les jours
2. Une ou plusieurs fois par semaine
3. Plus rarement

b. les variables non ordonnées

Exemple : parmi les distractions possibles le samedi soir, la personne interrogée préfère :

1. La télévision
2. Le théâtre
3. Le cinéma

Dans cette note, on traitera essentiellement des variables à deux modalités, que l'on notera 0 et 1. Quelques indications seront données sur les autres cas, qui correspondent aux réponses à plusieurs modalités.

II Pourquoi des modèles particuliers ?

On ne peut pas utiliser la même méthode que dans le cas continu puisqu'en particulier, la variable expliquée Y ne prenant que deux valeurs, la perturbation u suivrait obligatoirement une loi discrète, ce qui est incompatible avec les hypothèses habituelles de continuité et de normalité des résidus (voir Gouriéroux, 1989).

En effet, si on écrivait :

	$Y_i = X_i b + u_i$	pour l'individu i
alors on aurait:	$u_i = 1 - X_i \beta$	avec la probabilité p_i
	$u_i = -X_i \beta$	avec la probabilité $1 - p_i$
où	$p_i = P[Y_i = 1]$	soit une loi discrète pour u_i

III Niveau d'utilité, variables latentes

Les méthodes utilisées partent du principe que le phénomène observé est la manifestation visible d'une **variable latente** Z inobservable qui, elle, est continue. On se ramène alors conceptuellement à un modèle **d'analyse de la variance** sur cette variable latente, le problème à résoudre étant celui de l'estimation de ce modèle.

Exemple de cette variable latente : dans le cas de la possession d'un bien durable, la variable latente peut être « l'intensité du désir » de posséder le bien : tant que cette intensité reste inférieure à un certain seuil, on observe $Y_i = 0$ (le ménage i ne possède pas le bien), quand elle le dépasse on observe $Y_i = 1$ (le ménage i possède le bien). On peut aussi formuler le problème en terme de fonction d'utilité : pour le ménage i de caractéristiques X_i (âge, sexe de la personne de référence, revenu etc.), la possession du bien procure un **niveau d'utilité** $U(1, X_i)$, alors que la non possession procure un niveau $U(0, X_i)$.

On a alors :

$$Y_i = 1 \Leftrightarrow U(1, X_i) > U(0, X_i)$$

et $$Y_i = 0 \Leftrightarrow U(0, X_i) > U(1, X_i)$$

le ménage choisissant la situation qui lui procure le plus haut niveau d'utilité.

On se ramène au cas de la variable latente en posant :

$$Z_i = U(1, X_i) - U(0, X_i)$$

On a alors :

$$Y_i = 1 \Leftrightarrow Z_i > 0$$

et $$Y_i = 0 \Leftrightarrow Z_i < 0$$

Il y a possession du bien lorsque la variable latente Z_i dépasse le seuil 0.

IV Le (s) modèle (s) théorique (s)

Notons Y la variable dichotomique à expliquer, dite aussi variable dépendante, dont on supposera qu'elle prend les valeurs 0 et 1.

On observe les valeurs que prend Y sur un ensemble d'individus (ou de ménages) indicés par i, i = 1, ..., I. I est la taille de l'échantillon. Soit Z la **variable latente** sous-jacente au phénomène.

Le modèle postule une relation du type :

$$Z = Xb + u$$

où X est un ensemble de variables dites exogènes ou explicatives, qui peuvent être :

- des variables continues: le revenu, l'âge (dont l'effet est alors linéaire, voir plus loin dans les spécifications du modèle)

- des variables « discrétisées » : le revenu en tranches, l'âge décennal (ce qui permet de mettre en évidence des effets non linéaires)

- des variables qualitatives : la CSP, la catégorie de commune

Dans le cas de variables discrétisées ou qualitatives, il convient de choisir une situation de référence (voir ci-après).

La probabilité que l'individu i soit dans l'état $Y_i = 1$ est alors :

$$\begin{aligned} p_i = P[Y_i = 1] &= P[Z_i > 0] \\ &= P[X_i\beta > -u] \\ &= F(X_i\beta) \end{aligned}$$

si on note F la **fonction de répartition** de $-u$, c'est-à-dire la fonction définie par : $F(w) = P[-u < w]$.

Le choix du modèle porte sur le choix de F. Deux fonctions sont couramment utilisées et seront traitées ici :

- F = fonction de répartition de la loi normale (modèle PROBIT)
- F = fonction de répartition de la loi logistique (modèle LOGIT)

Toutefois, d'autres fonctions peuvent être choisies. Ainsi, la procédure `LOGISTIC` de SAS, dont on traitera plus loin, permet également de prendre pour F la fonction de répartition de la loi de Gompertz.

V Les modèles PROBIT et LOGIT

1. Le modèle PROBIT est celui pour lequel F est la fonction de répartition de la loi normale centrée réduite :

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt$$

ce qui donne:

$$P[Y = 1] = \Phi(X\beta) = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) dt$$

2. Le modèle LOGIT est celui pour lequel F est la fonction de répartition de la loi logistique:

$$F(w) = L(w) = \frac{\exp(w)}{1+\exp(w)} = \frac{1}{1+\exp(-w)}$$

ce qui donne :

$$P[Y = 1] = L(X\beta) = \frac{1}{1+\exp(-X\beta)}$$

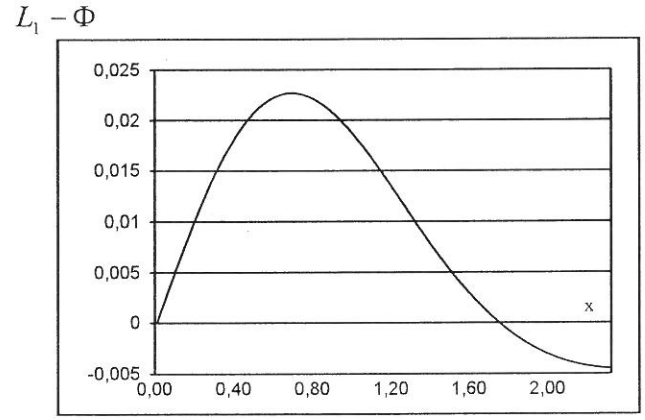
3. Comparaison des deux modèles

L (fonction de répartition de la loi logistique) et Φ (fonction de répartition de la loi normale) sont toutes les deux symétriques par rapport au point (0,1/2), et comprises entre 0 et 1 (ce qui convient pour représenter une probabilité).

La loi logistique de fonction de répartition L a pour moyenne 0, pour variance $\pi^2 / 3$; il est donc naturel de comparer à $\Phi(w)$, fonction de répartition de $N(0,1)$, la fonction $L_1(w)$ où

$$L_1(w) = \frac{1}{1+\exp(-\pi w / \sqrt{3})}$$

La figure ci-dessous donne en fonction de x, la différence $L_1(x) - \Phi(x)$ des fonctions de répartition :



(référence : Gouriéroux [1989]).

Ces lois étant proches, dans la plupart des cas pratiques on peut indifféremment choisir l'un ou l'autre modèle. Le modèle LOGIT a l'avantage d'une plus grande simplicité numérique, le modèle PROBIT est en revanche plus proche du modèle habituel de régression par les moindres carrés ordinaires.

Attention toutefois lorsque vous voudrez comparer les estimateurs obtenus à partir des différents modèles. La `Proc Logistic` utilise Φ et L (non pas L_1) : les estimateurs obtenus avec le modèle LOGIT seront donc $\pi/\sqrt{3}$ fois plus grands environ que ceux obtenus par le modèle PROBIT.

VI L'estimation : formules, précisions techniques

1. le principe de la méthode

La méthode d'estimation adoptée est celle du maximum de vraisemblance. L'enquête fournit I observations indépendantes (Y_i, X_i). Les Y_i sont des variables de Bernoulli $(1, p_i)$ où :

$$p_i = P[Y_i = 1]$$

La vraisemblance s'écrit alors :

- pour une observation : $p_i^{Y_i}(1-p_i)^{1-Y_i} = l_i(\beta)$
- pour I observations : $\Lambda_I(\beta) = \prod_{i=1}^{I} l_i(\beta)$

soit :

$$\Lambda_I(\beta) = \prod_{i=1}^{I} [F(X_i\beta)]^{Y_i} [1 - F(X_i\beta)]^{1-Y_i}$$

La log-vraisemblance s'écrit :

$$L_I(\beta) = \log(\Lambda_I(\beta)) = \sum_{i=1}^{I} \log(l_i(\beta))$$

soit :

$$l = L_I(\beta) = \sum_{i=1}^{I} Y_i \log(F(X_i\beta)) + \sum_{i=1}^{I} (1 - Y_i) \log(1 - F(X_i\beta))$$

On en déduit la dérivée :

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{I} \frac{Y_i - F(X_i\beta)}{F(X_i\beta)(1 - F(X_i\beta))} f(X_i\beta) X_i$$

où f est la dérivée de F, et la matrice des dérivées secondes, ou *Hessien* :

$$\frac{\partial^2 l}{\partial\beta\partial\beta'} = -\sum_{i=1}^{I} \left[\frac{Y_i}{F^2(X_i\beta)} + \frac{1 - Y_i}{(1 - F(X_i\beta))^2} \right] f^2(X_i\beta) X_i X_i' + \sum_{i=1}^{I} \frac{Y_i - F(X_i\beta)}{F(X_i\beta)(1 - F(X_i\beta))} f'(X_i\beta) X_i X_i'$$

ainsi que la matrice d'information de *Fisher* :

$$I_F(\beta) = -\left(E \frac{\partial^2 l}{\partial\beta\partial\beta'}\right)$$

La procédure d'estimation consiste à rechercher la valeur $\hat{\beta}$ de β qui maximise la vraisemblance ou plus précisément son logarithme $L_I(\beta)$, noté l .

2. L'algorithme utilisé

Dans le cas des modèles LOGIT ou PROBIT, on montre aisément que la log-vraisemblance l est concave. $\hat{\beta}$ est alors la solution de l'équation :

$$\frac{\partial l}{\partial \beta} = 0$$

C'est-à-dire :

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{I} \frac{Y_i - F(X_i\beta)}{F(X_i\beta)(1-F(X_i\beta))} f(X_i\beta)X_i = 0$$

Cette solution est unique dans les cas usuels de non-dégénérescence. Donc toute procédure itérative convergente (dont l'emploi pour résoudre l'équation différentielle est nécessaire car l'équation est non-linéaire) converge vers $\hat{\beta}$. La procédure employée dans la plupart des cas est basée sur l'algorithme de Newton-Raphson.

Dans le cas des GLM, on utilise souvent une autre procédure, celle de l'algorithme de Fisher (*Fisher scoring*). Cet algorithme ressemble à celui de Newton-Raphson, la différence étant que le Fisher scoring utilise **l'espérance** de la matrice des dérivées secondes au lieu de la matrice elle-même.

Soit $\beta^{(k)}$ la k-ième approximation pour l'EMV $\hat{\beta}$. Dans la méthode de Newton-Raphson, on a :

$$\beta^{(k+1)} = \beta^{(k)} - (H^{(k)})^{-1} q^{(k)}$$

où H est la matrice hessienne ayant pour éléments $\frac{\partial^2 L(\beta)}{\partial\beta_h \partial\beta_i}$, q est le vecteur des dérivées ayant pour éléments $\frac{\partial L(\beta)}{\partial\beta_j}$; $H^{(k)}$ et $q^{(k)}$ sont évaluées en $\beta = \beta^{(k)}$.

La formule du *Fisher scoring* s'écrit :

$$\beta^{(k+1)} = \beta^{(k)} + (I_F(\beta^{(k)}))^{-1} q^{(k)}$$

où $I_F(\beta^{(k)})$ est la k-ième approximation de la matrice d'information de *Fisher* estimée. Autrement dit, $I_F(\beta^{(k)})$ a pour éléments $-(E\frac{\partial^2 L(\beta)}{\partial\beta_h \partial\beta_i})$, évaluée en $\beta = \beta^{(k)}$.

On montre que la méthode du *Fisher scoring* peut s'interpréter comme une succession de moindres carrés, pondérés par des poids qui changent à chaque itération. L'estimation de la matrice de variance-covariance est un sous-produit de la méthode. Pour cette raison, l'algorithme employé est appelé « moindres carrés repondérés itératifs » (Iteratively Reweighted Least Squares ou IRLS).

La procédure employée par la `Proc Logistic` de SAS utilise cette méthode itérative de moindres carrés repondérés. A partir d'une valeur initiale $\hat{\beta}^{(0)}$, on corrige l'estimation selon une formule du type :

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + c^{(i)}$$

jusqu'à obtenir la stabilité, en l'occurrence jusqu'au moment où la valeur absolue de la différence entre les valeurs calculées pour le logarithme de la vraisemblance à deux étapes successives soit en deçà d'un seuil fixé à l'avance. Pour la `Proc Logistic`, toutefois, on considère que les itérations ont convergé lorsque la différence maximale entre les estimateurs des différents paramètres est inférieure à un seuil, par défaut 10^{-4}.

Pour plus de détails sur la méthode IRLS voir SAS/STAT User's guide, vol.2. Pour plus de détails sur la méthode de Newton-Raphson, voir AGRESTI [1990] ou GOURIEROUX [1989].

3. Quelques propriétés asymptotiques de l'estimateur du maximum de vraisemblance

Sous des hypothèses très générales, l'estimateur du maximum de vraisemblance a de bonnes propriétés. Il est asymptotiquement (i.e. lorsque I est grand) normal :

$$\sqrt{I}(\hat{\beta}-\beta) \xrightarrow{asymp.} N(0, I_F^{-1}(\beta))$$

où $I_F(\beta)$ désigne la matrice d'information de Fisher. La matrice de variance-covariance asymptotique de l'estimateur du maximum de vraisemblance s'écrit donc :

$$V\hat{\beta} = -(E\frac{\partial^2 l}{\partial\beta\partial\beta'})^{-1}$$

Or , conditionnellement aux X_i, on a :

$$E(\frac{\partial^2 l}{\partial\beta\partial\beta'}/X_i) = -\sum_{i=1}^{I}\frac{f^2(X_i\beta)}{F(X_i\beta)(1-F(X_i\beta))}X_iX_i'$$

La matrice de variance-covariance asymptotique (conditionnelle) de $\hat{\beta}$ vaut donc :

$$V\hat{\beta} = \left[\sum_{i=1}^{I}\frac{f^2(X_i\beta)}{F(X_i\beta)(1-F(X_i\beta))}X_iX_i'\right]^{-1}$$

On en obtient un estimateur en calculant la valeur précédente au point $\hat{\beta}$.

Cas particulier du modèle LOGIT

Dans ce cas, on a :

$$F(w) = \frac{1}{1+\exp(-w)}$$

$$f(w) = F'(w) = \frac{\exp(-w)}{(1+\exp(-w))^2} = \frac{1}{1+\exp(-w)}\frac{\exp(-w)}{1+\exp(-w)}$$

$$f(w) = F(w)(1-F(w))$$

avec :

$$p_i = P[Y_i = 1] = \frac{1}{1+\exp(-X_i\beta)}$$

$$l = \log(\Lambda_I(\beta)) = \sum_{i=1}^{I}(1-Y_i)(-X_i\beta) - \sum_{i=1}^{I}\log(1+\exp(-X_i\beta))$$

$$\frac{\partial l}{\partial\beta} = \sum_{i=1}^{I}\frac{Y_i - F(X_i\beta)}{F(X_i\beta)(1-F(X_i\beta))}f(X_i\beta)X_i = \sum_{i=1}^{I}(Y_i - F(X_i\beta))X_i$$

et donc la dérivée seconde de la log-vraisemblance se simplifie en :

$$\frac{\partial^2 l}{\partial\beta\partial\beta'} = -\sum_{i=1}^{I} f(X_i\beta)X_iX_i' = -\sum_{i=1}^{I} F(X_i\beta)(1-F(X_i\beta))X_iX_i' = -\sum_{i=1}^{I} p_i(1-p_i))X_iX_i'$$

Il en résulte que le *Hessien* ne dépend pas des observations de Y_i. Il est alors égal à son espérance conditionnelle aux X_i. De ce fait, le *Fisher scoring* et la méthode de Newton-Raphson sont équivalents. En particulier, l'estimateur de la matrice de variance-covariance de $\hat{\beta}$ peut s'écrire:

$$\hat{V}\hat{\beta} = \left[\sum_{i=1}^{I} X_i'X_i\hat{p}_i(1-\hat{p}_i)\right]^{-1}$$

où :

$$\hat{p}_i = \frac{1}{1+\exp(-X_i\hat{\beta})}$$

représente l'estimation de la probabilité de choix (par exemple de choisir de posséder un bien) pour l'individu i de caractéristiques individuelles X_i.

Cas particulier du modèle PROBIT

f est la densité de la loi normale centrée réduite et F son intégrale.

VII Les tests

1. Test de la nullité d'un coefficient

On veut tester la nullité du coefficient β_j, c'est à dire de la j^{ème} composante du vecteur de paramètres β. β_j est le coefficient correspondant à la j^{ème} variable explicative (jème colonne de la matrice X).

On considère la statistique de Student :

$$\frac{\hat{\beta}_j}{\sqrt{\hat{V}\hat{\beta}_j}}$$ où

- $\hat{\beta}_j$ est la j^{ème} composante de l'estimateur
- $\hat{V}\hat{\beta}_j$ est le j^{ème} coefficient de la diagonale de la matrice de variance-covariance estimée de $\hat{\beta}$
- $\sqrt{\hat{V}\hat{\beta}_j}$ en est l'écart-type estimé (*standard deviation*)

On compare habituellement cette statistique au seuil de significativité à 5% d'une loi normale (environ 2).

Dans la procédure Logistic de SAS, la significativité de chaque coefficient $\hat{\beta}_j$ est testée à partir de la statistique de Wald :

$$W = \frac{\hat{\beta}_j^2}{\hat{V}\hat{\beta}_j}$$

soit le carré de la statistique de Student.

Cette statistique suit asymptotiquement une loi du χ^2 à 1 degré de liberté. l'hypothèse de la nullité de $\hat{\beta}_j$ est rejetée lorsque la statistique de Wald dépasse un certain seuil, environ 4 pour une significativité à 5 %.

2. Test d'une liaison de la forme $\sum_{k=1}^{I} \lambda_k \beta_k = C$

Si on note $\hat{V}\hat{\beta}$ la matrice de variance-covariance estimée de l'estimateur $\hat{\beta}$ et Q' le vecteur ligne $(\lambda_1,...,\lambda_I)$, on a le résultat asymptotique suivant :

$$\frac{Q'\hat{\beta} - C}{\sqrt{Q'(\hat{V}\hat{\beta})Q}} \xrightarrow{asymp.} N(0,1)$$

si l'hypothèse $Q'\beta = C$ est vraie

Si l'hypothèse alternative du test est $Q'\beta \neq C$, l'hypothèse « nulle » est rejetée si la valeur absolue de la statistique précédente dépasse un certain seuil de significativité.

Le cas 1 est bien sûr un cas particulier de 2, lorsque seul λ_j est non nul et $C = 0$.

3. Test de la nullité d'un ensemble de coefficients

On peut souhaiter tester la nullité d'un ensemble de q coefficients (par exemple tous ceux concernant les différentes variables introduites pour représenter une dimension explicative (cf infra) telle que la CSP, ou bien le revenu en tranches, ou bien l'âge quinquennal etc.). On peut souhaiter tester également la nullité de l'ensemble des coefficients.

L'hypothèse de la nullité d'un ensemble de q coefficients s'écrit sous la forme $Q'\beta = 0$, où Q' est une matrice diagonale où seuls les coefficients correspondant aux β_j dont on veut tester la nullité sont égaux à 1, les autres étants nuls. Par exemple, dans le cas où:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

et où on veut tester $\beta_1 = 0$ et $\beta_2 = 0$, on aura:

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

On dispose de plusieurs tests, par exemple :

- le test de Wald
- le test du rapport de vraisemblance

*** Test de Wald**

$$W = (Q'\hat{\beta})' \left[Q' (\hat{V}\hat{\beta}) Q \right]^{-1} (Q'\hat{\beta}) \xrightarrow{asymp.} \chi^2_q$$

W tend asymptotiquement vers un χ^2 à q degrés de liberté. Rappelons que dans le cas d'une variable catégorielle à p modalités, comme les CSP, l'une des modalités est prise comme niveau de référence et son coefficient est donc nul. La statistique de Wald sur les coefficients des modalités qui restent sera donc convergente asymptotiquement vers un χ^2 à p-1 degrés de liberté.

Ici encore, l'hypothèse « nulle » $Q'\beta = 0$ sera rejetée lorsque la valeur de la statistique de Wald dépassera un seuil critique.

* **Test du rapport de vraisemblance**

Si l désigne la log-vraisemblance, $\hat{\beta}$ l'estimateur du maximum de vraisemblance, $\hat{\beta}_0$ l'estimateur du maximum de vraisemblance sous la contrainte $Q'\beta = 0$ on a :

$$LRT = 2(l(\hat{\beta}) - l(\hat{\beta}_0)) \xrightarrow{asymp.} \chi^2_q$$

Ici aussi, l'hypothèse de nullité simultanée des coefficients considérés doit être rejetée si la valeur de la statistique dépasse un seuil critique.

Application : choix entre 2 modèles dont l'un est une version « réduite » de l'autre.

Modèle 1 : les variables explicatives sont $X_1,...,X_p$

Modèle 2 : $X_1,...,X_p,X_{p+1},...,X_{p+k}$

Préférer 1 à 2, c'est accepter l'hypothèse que, dans le second modèle, les k coefficients $\beta_{p+1},\ldots,\beta_{p+k}$ sont nuls. Cette hypothèse s'écrit sous la forme $Q'\beta = 0$ comme on l'a déjà vu.

On choisira le modèle 2 si :

$$LRT = 2(l(\hat{\beta}) - l(\hat{\beta}_0))$$

est supérieur à la valeur critique au seuil de a % du χ^2 à k degrés de liberté.

* **Attention** : ce type de choix entre 2 modèles dont l'un est une version réduite de l'autre se présente en particulier dans le cas d'estimations BACKWARD (on retire des variables au modèle selon certains critères de choix), FORWARD (on en ajoute), ou STEPWISE (alternativement, on retire et on ajoute des variables au modèle). Toutefois, la `Proc Logistic` de SAS choisit entre les modèles en utilisant la **statistique du score pour la procédure FORWARD et la statistique de Wald pour la procédure BACKWARD.**.

La statistique du score est une forme quadratique construite à partir du vecteur des dérivées partielles de la log-vraisemblance par rapport au vecteur de paramètres β, et évaluée en β_0 (c'est à dire sous l'hypothèse nulle). On a alors :

$$S = (\frac{\partial l}{\partial \beta}(\beta_0))' I_F(\beta_0)^{-1} \frac{\partial l}{\partial \beta}(\beta_0)$$

(où $I_F(\beta)$ est la matrice d'information de Fischer), qui suit asymptotiquement une loi du χ^2 à k degrés de liberté.

On choisira alors le modèle 2 (c'est à dire celui qui comporte le plus de variables explicatives) lorsque S sera supérieur à la valeur critique au seuil de a % du χ^2 à k degrés de liberté. SAS édite la « p-value » de la statistique S dite aussi « statistique du χ^2 résiduel », c'est à dire la probabilité que sous l'hypothèse nulle

(modèle 1) la statistique S dépasse la valeur observée. Cette « p-value» doit être faible pour choisir le modèle 2.

4. Cas plus général : test d'une hypothèse linéaire de la forme $Q'\beta = C$

où : Q' est une matrice de coefficients constants connus de dimension q x K (K nombre de variables dans le modèle estimé, y compris la constante), et C est un vecteur de constantes connues, déterminées par l'utilisateur.

Les q lignes de Q sont linéairement indépendantes. On voit que les cas traités précédemment sont tous des cas particuliers de celui-ci.

On peut ici encore utiliser le test de Wald :

$$W = (Q'\hat{\beta} - C)'\left[Q'(\hat{V}\hat{\beta})Q\right]^{-1}(Q'\hat{\beta} - C) \xrightarrow{asymp.} \chi^2_q$$

ou celui du rapport de vraisemblance :

$$LRT = 2(l(\hat{\beta}) - l(\hat{\beta}_0)) \xrightarrow{asymp.} \chi^2_q$$

où $\hat{\beta}_0$ est l'estimateur obtenu en maximisant la vraisemblance sous la contrainte $Q'\beta = C$

Comme précédemment, l'hypothèse $Q'\beta = C$ doit être rejetée si la valeur de la statistique dépasse un certain seuil.

5. Tests de la validité générale du modèle

Existe-t-il des statistiques permettant de juger de la bonne adéquation du modèle, en jouant un rôle analogue à celui du R^2 classique ? Les auteurs en ont proposé plusieurs, souvent critiquables à un titre ou à un autre. Il est en particulier difficile d'apporter les corrections adéquates pour comparer des modèles ayant des nombres de degrés de liberté différents.

Voici celles fournies par la `Proc Logistic` :

- Le rapport de vraisemblance (l'hypothèse nulle étant celle où le modèle contient la seule constante) ;

- la statistique du score (ou du χ^2 résiduel) déjà définie plus haut ;

- le critère d'Akaike

$$AIC = -2\log L + 2K$$

où K est le nombre de paramètres à estimer ;

- le critère de Schwartz

$$SC = -2\log L + K\log I$$

où I est le nombre total d'observations.

Les critères de Schwartz et d'Akaike sont utiles pour comparer des modèles différents portant sur les mêmes données. On préférera le modèle pour lequel ces statistiques ont la valeur la plus faible.

D'autres approches permettent d'évaluer la capacité prédictive du modèle :

• Les prédictions et observations concordantes

On considère toutes les paires d'observations ayant des valeurs observées de Y différentes, soient 1 et 0. Parmi ces paires, on compte celles pour lesquelles l'observation où $Y = 1$ a une probabilité estimée que $Y = 1$ plus grande que l'observation où $Y = 0$. On dit alors que la paire est concordante. Elle est discordante lorsque la probabilité que $Y = 1$ est plus faible pour l'observation où $Y = 1$ que pour celle où $Y = 0$. Les paires qui ne sont ni concordantes ni discordantes sont dites « liées » (tied) ou « ex-aequo ». Si I est le nombre total d'observations, t le nombre de paires ayant des valeurs observées de Y différentes, n_c le nombre de paires concordantes, n_d le nombre des paires discordantes, $t - n_c - n_d$ le nombre de paires « liées », SAS calcule quatre indices de « corrélation du rang » (rank correlation) :

C	$= (n_c + 0{,}5(t - n_c - n_d))/t$
Somer's D	$= (n_c - n_d)/t$
Goodman - Kruskal Gamma	$= (n_c - n_d)/(n_c + n_d)$
Kendall's Tau - a	$= (n_c - n_d)/(0{,}5I(I-1))$

Ces quatre indices sont en quelque sorte des mesures d'association entre la probabilité prédite et la valeur de la variable explicative. Cette association est d'autant plus forte (et on est d'autant plus satisfait) que les indices sont élevés, c'est-à-dire proches de 1. En effet tous ces indices sont croissants lorsque n_c croît, décroissants lorsque n_d croît et varient entre les bornes suivantes :

C :	entre 0 et 1
Somer's D :	entre - 1 et + 1
Gamma :	entre - 1 et + 1
Kendall's Tau-a :	entre - 1 et + 1

Le cas extrême où l'indice prend la valeur + 1 est celui où la totalité des paires ayant pour un élément $Y = 0$ et pour l'autre $Y = 1$ sont concordantes (c'est-à-dire que la probabilité estimée que $Y = 1$ est plus forte pour l'observation telle que $Y = 1$) : la prévision correspond « au mieux » à la réalité.

• les tables de classification (voir l'option `Ctable Pprob=` dans l'instruction `Model`)

L'idée de base de ces tables est de « prédire » Y_i par $\hat{Y}_i$ de la façon suivante :

$\hat{Y}_i = 1$	si la probabilité estimée de valoir 1 dépasse un certain seuil
$\hat{Y}_i = 0$	sinon.

SAS fait varier le seuil de 0 à 1 et donne, pour chaque valeur, quatre ratios définis comme suit :

La « sensibilité » (sensitivity) est la proportion de vraies valeurs 1 qui sont prédites valoir 1.
La « spécificité » (specificity) répond à la définition analogue pour les valeurs 0.
Le « taux d'erreur par excès » (false positive rate) est la proportion de prédictions 1 qui valent en réalité 0.
Le « taux d'erreur par défaut » (false negative rate) la proportion de prédictions 0 qui valent en réalité 1.

Lorsque le seuil est très bas, la prédiction vaut toujours 1. Le modèle ne se trompe jamais pour prédire l'événement, mais il se trompe toujours pour prédire le non événement. La sensibilité est de 100, et la spécificité de 0. Le taux d'erreur par excès correspond à la fréquence des observations pour lesquelles $Y = 0$. A l'opposé, lorsque le seuil est trop élevé, la prédiction est toujours de 0. La sensibilité est nulle, et la spécificité de 100. Le taux d'erreur par défaut correspond alors à la fréquence des observations pour lesquelles $Y = 1$. Les seuils compris entre ces deux extrêmes balaient la plage des situations intermédiaires.

Il faut se garder d'utiliser cette table de classification pour juger de la capacité prédictive du modèle. En effet, si le modèle est vrai dans la population, quel que soit le seuil α choisi, **il y aura toujours des observations pour lesquelles** $F(X_i\beta) > \alpha$ **et** $Y_i = 0$. Autrement dit, le critère implicite de « bon modèle » qui est derrière les tables de classifications, à savoir « si le modèle était parfait, tout serait prédit parfaitement » n'est qu'une illusion.

Exemple : supposons la population composée de deux groupes de 1000 individus ayant les mêmes X_i mais pas nécessairement les mêmes u_i, de sorte que certains recourent à la pratique et d'autres non. La population est telle que $F(X_i\beta) = 0.8$ pour les individus du premier groupe, et $F(X_i\beta) = 0.2$ pour les individus du deuxième groupe. Ce que le modèle dit, c'est qu'environ 800 personnes du premier groupe et environ 200 personnes du deuxième groupe auront $Y = 1$. Or, la construction de la `CTABLE` conduit à calculer $F(X_i\hat{\beta})$, qui a la même valeur pour tous les individus du même groupe. Quelque soit le seuil, la procédure affecte donc la même valeur à tous des individus de chaque groupe. Ceci n'est pas cohérent avec l'hypothèse aléatoire justifiant le modèle.

Pour formaliser ceci, notons que si le modèle est vrai, le nombre d'observations valant 1 dans l'échantillon est, sous ce modèle:

$$C = \sum_{i=1}^{I} \mathrm{I}\{Y_i = 1\}.$$

C'est une variable aléatoire dont on observe la réalisation. Son espérance et sa variance sont :

$$E(C) = \sum_{i=1}^{I} F(X_i\beta), \quad V(C) = \sum_{i=1}^{I} p_i(1-p_i).$$

Un estimateur « raisonnable » de C serait :

$$\hat{C} = \sum_{i=1}^{I} F(X_i\hat{\beta}).$$

Or, les quantités décrites dans les tables de classificiation proviennent d'estimateurs asymptotiquement du type $\tilde{C} = \sum_{i=1}^{I} \mathrm{I}\{F(X_i\hat{\beta}) > \alpha\}$. Dans le cas où $\alpha = 1/2$, $\tilde{C} = \sum_{i=1}^{I} \mathrm{I}\{X_i\hat{\beta} > 0\}$, qui diffère de $\hat{C}$. Plus généralement, et quelque soit α, ce type de validation de modèle ne s'applique guère aux problèmes de nature statistique. Il est davantage destiné aux applications médicales, qui cherchent à contrôler le niveau ou la puissance des tests.

• Les ODD RATIOS

Lors de l'estimation des modèles Logit (et d'eux seulement), la `Proc Logistic` fournit à côté de chaque coefficient une statistique d'ODD RATIO, qui présente certaines propriétés. Celles-ci proviennent de la généralisation de pratiques à l'origine destinées à l'analyse de tableaux croisés.

Analyse de tableaux croisés

Admettons que l'on souhaite répondre à la question : les femmes font-elles plus de sport que les hommes ? Comment décrire l'écart entre les deux pratiques ?

On dispose de deux variables :

- une variable expliquée Y valant 1 pour les individus faisant du sport, 0 pour les autres.
- une variable explicative X, valant 0 pour les hommes et 1 pour les femmes.

On suppose que les femmes sont en proportion 54 % dans la population, que 30 % des hommes et 50 % des femmes font du sport ; la proportion de personnes faisant du sport dans la population est donc de 40,8%.
Pour comparer les hommes et les femmes, on définit :

a) le <u>risque relatif</u> :

$$r = \frac{\text{Probabilit é de faire du sport quand on est un homme}}{\text{probabilit é de faire du sport quand on est une femme}} = \frac{30}{50} = 0{,}6$$

b) la <u>différence des taux de pratique</u> au sein des deux sous-populations :

d=50 %-30 %=20 points en faveur des femmes.

c) les <u>chances</u> (ODDS) pour chaque sexe.

$$\Omega = \frac{\text{Probabilité de faire du sport}}{\text{probabilité de ne pas faire de sport}} = \frac{p}{1-p}$$

Pour se représenter la signification des chances on peut imaginer que l'on prend un pari sur le fait de faire du sport ou pas. Si l'événement « ne pas faire de sport » est par hypothèse à la cote 1, alors « faire du sport » est à la cote ODDS

On peut calculer autant de ODDS que de modalités de la variable explicative.

Pour les hommes $\Omega_H = \frac{n_{11}}{n_{12}} = \frac{30}{70} = 0.43$

Pour les femmes $\Omega_F = \frac{n_{21}}{n_{22}} = \frac{50}{50} = 1$

Le logarithme de la ième chance est appelé « le ième logit » (i=H ou F ici).

d) Enfin, on définit le <u>rapport des chances (ODDS RATIO)</u>

$OR_{H/F} = \frac{\Omega_H}{\Omega_F} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$ Ici $OR_{H/F} = \frac{0.43}{1} = 0.43$

Si on prend comme catégorie de référence les hommes au lieu des femmes, le ODDS RATIO est transformé en son inverse.

$OR_{H/F} = \frac{1}{OR_{F/H}}$ et donc $\log(OR_{H/F}) = -\log(OR_{F/H})$

Le ODDS RATIO a comme propriété d'être inchangé si l'on multiplie tous les effectifs d'une ligne ou d'une colonne par une constante strictement positive.

	Sport	Pas de sport	total

hommes	n_{11}	n_{12}	n_{1+}	x C
femmes	n_{21}	n_{22}	n_{2+}	x D
total	n_{+1}	n_{+2}	**n**	
	x A	x B		

Le nouveau ODDS RATIO vaut alors :

$$OR_{H/F} = \frac{\Omega_H}{\Omega_F} = \frac{ACn_{11}BDn_{22}}{ADn_{21}BCn_{12}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

En outre, le ODDS RATIO ne dépend pas des marges : il est robuste par rapport à la représentativité de l'échantillon.

Supposons que pour connaître ces chiffres, on décide de réaliser une enquête. Pour la commodité du raisonnement, on suppose que l'on peut identifier exactement les sportifs et les non sportifs (par exemple, on décide qu'est sportif celui qui possède une licence dans un sport, ce qui permet d'avoir une base de sondage). On suppose également que les enquêtes donnent toutes les proportions exactes.

Première enquête

On tire 5000 personnes dans la population, et on les classe suivant le sexe et la pratique du sport (échantillonnage binomial)

	Sport	Pas de sport	total
hommes	690 (n_{11})	1610(n_{12})	2300 (n_{1+})
femmes	1350 (n_{21})	1350 (n_{22})	2700 (n_{2+})
total	2040 (n_{+1})	2960 (n_{+2})	**5000 (n, fixé)**

A partir de cette enquête, on peut estimer toutes les proportions désirées. Par exemple en ligne

	Sport	Pas de sport	total
hommes	30 %	70 %	100 %
femmes	50 %	50 %	100 %
total	40,8 %	59,2 %	100 %

On vérifie que tous les indicateurs de disparité entre hommes et femmes que nous avons considérés sont égaux à leur vraie valeur dans la population.

Supposons par exemple que notre enquête soit maintenant menée de la manière suivante :

Deuxième enquête :

On enquête 1000 sportifs et 1000 non sportifs (on conditionne l'échantillonnage sur la variable expliquée). Le tableau croisant le sexe et la pratique du sport devient :

	Sport	Pas de sport	total
hommes	338	544	882
femmes	662	456	1118
	1000=n_{+1} fixé	**1000=n_{+2} fixé**	**2000**

A partir de cette enquête, on ne peut pas estimer correctement la part des hommes ou des femmes qui font du sport. On obtiendrait : 338 / 882=38,3 % pour les hommes

662 / 1118=59,2 % pour les femmes.

On ne peut plus estimer correctement le risque relatif :

$$r = \frac{n_{11} n_{2+}}{n_{1+} n_{21}} = \frac{338.1118}{882.662} = 0,65$$

ni la différence des taux de pratique au sein des deux sous-populations = 20,9 points

ni les ODDS pour les hommes et les femmes

$$\Omega_H = \frac{n_{11}}{n_{12}} = \frac{338}{544} = 0,62$$

En revanche, on peut toujours estimer correctement le ODDS RATIO (OR). Celui-ci est égal à :

$$OR_{H/F} = \frac{\Omega_H}{\Omega_F} = \frac{n_{11} n_{22}}{n_{21} n_{12}} = \frac{338.456}{544.662} = 0,43$$

Le ODDS RATIO est donc adapté à différents modes d'échantillonnage :
- l'échantillonnage binomial (celui des enquêtes ménages à l'INSEE, en général)
- celui où on tire indépendamment dans les catégories de la variable explicative (appelé « tirage binomial indépendant »)
- celui où on tire indépendamment dans les catégories de la variable expliquée (appelé « étude rétrospective », terme provenant du vocabulaire médical, car les études de maladies adoptent souvent ce mode d'échantillonnage).

Généralisation à plusieurs variables

S'il y a plusieurs variables explicatives (sexe, tranche d'âge, CSP), on peut encore définir les ODDS RATIO, par rapport à une situation de référence (une « case » du tableau croisé de toutes les variables explicatives). Par exemple : être un homme, de plus de 40 ans, employé.

La chance (ODDS) dans la catégorie de référence est alors :

$$\Omega_R = \frac{\text{nombre de personnes qui font du sport dans cette caté gorie}}{\text{nombre de personnes qui ne font pas de sport dans cette catégorie}}$$

Si maintenant on s'intéresse à une catégorie *i* de la population (par exemple, Femme, de 30 à 40 ans, cadre), on peut définir la chance (ODDS) dans cette catégorie :

$$\Omega_i = \frac{\text{nombre de personnes qui font du sport dans cette caté gorie}}{\text{nombre de personnes qui ne font pas de sport dans cette catégorie}}$$

L' ODDS RATIO pour cette catégorie est défini par $OR_{i/R} = \frac{\Omega_i}{\Omega_R}$, ou, par définition du logit (défini, rappelons-le, comme le logarithme de la chance): $\log(OR_{i/R}) = \text{logit}(p_i) - \text{logit}(p_R)$. Selon les valeurs des probabilités de faire du sport dans chaque case du tableau croisé des variables explicatives, $\log(OR_{i/R})$ peut varier de moins l'infini à plus l'infini. On peut songer à le modéliser comme une fonction linéaire des différentes variables explicatives:

$$\log(OR_{i/R}) = \beta_1.(\text{appartenance à la tranche d'age i}) + \beta_2.(\text{appartenance à la CSP j}) + \beta_3.(\text{appartenance au sexe masculin ou fé minin})$$

Sous cette forme, les coefficients ne dépendent pas du mode d'échantillonnage, parmi les trois modes envisagés, car l'ODDS RATIO est le même dans les trois cas.

Le modèle s'écrit de manière équivalente

$$\boxed{\begin{aligned} &\text{logit}(p_i) = \text{logit}(p_R) + \sum_{i=1}^{p} \beta_i X_i \\ &\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{i=1}^{p} \beta_i X_i \\ &\text{ODDS RATIO} = \exp\left(\sum_{i=1}^{p} \beta_i X_i\right) \end{aligned}} \quad (1)$$

C'est le principe du modèle LOGIT. La troisième formule est celle que l'on retrouve en sortie de la `Proc Logistic` lorsque seul un X_i est non nul.

A partir de la formule (1), seule la constante (qui représente le logit de la catégorie de référence) est affectée par le mode d'échantillonnage. *(Pour un exposé théorique complet et intuitif sur l'analyse des tableaux croisés et les modélisations possibles, voir Agresti [1990]).*

VIII Mise en oeuvre de la procédure LOGISTIC de SAS

1. Quelques remarques et mises en garde préalables

• La procédure `Logistic` , décrite dans ce manuel, ajuste des modèles à résidus logistiques (LOGIT) ou normaux (PROBIT) ou encore correspondant à la loi de Gompertz (voir plus haut). Dans la procédure, la variable dépendante doit être soit dichotomique (ce qui est le cas traité dans cette note), soit polytomique ordonnée. La procédure Probit a sensiblement les mêmes propriétés que la procédure `Logistic`. Ses fonctionnalités, assez réduites dans les premières versions de SAS, sont aujourd'hui analogues. La procédure Probit présente cependant l'avantage, grâce à l'une de ses options, de pouvoir prendre en compte des comportements pour lesquels le taux de saturation dans la population sont inférieurs à 100 % *(pour en savoir plus, se référer à la brochure SAS)*. Les syntaxes étant légèrement différentes, le choix entre la `Proc Logistic` et la `Proc Probit` est d'abord affaire d'habitude.

La procédure `Catmod` traite elle aussi les modèles dichotomiques et polytomiques ordonnés. Mais elle est avant tout destinée à estimer les modèles polytomiques non ordonnés, ou les modèles de transition markoviens en temps discret. De ce fait, elle est plus complexe et les procédures `Logistic` ou `Probit` sont vivement conseillées pour les modèles dichotomiques et polytomiques ordonnés.

• La procédure `Logistic` traite les variables explicatives comme si elles étaient continues. Il convient donc de dichotomiser les variables explicatives qualitatives, telles que CSP, sexe, mais aussi tranches de revenu ou d'âge. La procédure `Probit` dispense d'une telle opération, car elle comporte une option `Class` comme la `Proc GLM`. Mais l'usage d'une telle option interdit de récupérer dans un data séparé les estimateurs correspondant aux variables explicatives. En outre, les choix par défaut de l'option `Class` (modalité de référence pouvant être vide ou presque,...) sont parfois gênants, et obligent à recourir à des recodifications parfois plus lourdes que la fabrication des variables muettes.

Il est maintenant utile de formuler quelques remarques générales concernant la dichotomisation des variables explicatives qualitatives.

↳ ***La variable explicative X est déjà à deux modalités, 0 et 1*** : on ne change rien. On fera figurer X dans la liste des variables explicatives et la procédure considérera que 0 est la modalité de référence.

↳ ***La variable explicative X est à deux modalités quelconques*** (par exemple : 8 et 9). si on choisit 8 pour modalité de référence, on fera figurer dans la liste des variables explicatives X1 défini au préalable par :

X1 = (X = 9) ;

↳ ***La variable explicative X a n modalités prenant les valeurs 1, ..., n.***

On utilisera l'instruction `Array`.
Exemple : la catégorie socio-professionnelle est la variable PPCS qui vaut de 1 à 8.
On écrira :

```
Array P (J)  PPCS1 - PPCS8;
Do J = 1 TO 8;
P = (PPCS = J);
End;
```

♆ La variable explicative X a n + 1 modalités prenant les valeurs 0, 1, ..., n.

Première solution : on recodifie au préalable par X=X+1 et on se ramène au cas précédent.
Deuxième solution : prenons l'exemple du diplôme de la personne de référence, DIPLOPR, qui varie de 0 à 5. On écrira

```
Array  D(M)    DIPR0-DIPR5 ;
Do  M = 1    To 6 ;
D = (DIPLOPR = M-1);
End;
```

♆ On veut à la fois dichotomiser et regrouper des modalités.

Exemple : le revenu du ménage est indiqué par la variable REVENU qui prend des valeurs de 1 à 8 (8 tranches). On veut opérer les regroupements suivants :

1 et 2, 3 et 4, 5 à 7, 8

On écrira :

```
REV1 = (REVENU = 1 ! REVENU = 2) ;
REV2 = (REVENU = 3 ! REVENU = 4) ;
REV3 = (5 < = REVENU < = 7) ;
REV4 = (REVENU = 8) ;
```

Ne pas oublier qu'il faut une modalité de référence (sinon `Proc Logistic` prend la dernière). Cette modalité est celle qui est omise dans la liste des variables de l'instruction `Model`.

♆ La procédure Logistic ajuste le modèle sur la probabilité de la modalité la plus faible.

Si donc vous avez codé :

$$Y = \begin{cases} 0 & \text{je ne possède pas un bien} \\ 1 & \text{je le possède} \end{cases}$$

SAS modélise la probabilité de ne pas avoir le bien. Les coefficients de la régression seront positifs pour les modalités explicatives correspondant à une plus forte probabilité de ne pas posséder le bien.

Les **vrais** coefficients modélisant la probabilité inverse sont tout simplement l'opposé des coefficients obtenus. En effet , si on pose $Z_i = 1 - Y_i$, on a pour le modèle LOGIT:

$$\begin{aligned} \Pr(Z_i = 0) &= \Pr(Y_i = 1) &&= \frac{1}{1+\exp(-X\beta_{(Y)})} \\ &= 1 - \Pr(Z_i = 1) &&= 1 - \frac{1}{1+\exp(-X\beta_{(Z)})} \quad = \frac{1}{1+\exp(X\beta_{(Z)})} \end{aligned}$$

Donc : $\beta_{(Y)} = -\beta_{(Z)}$.

Par le principe d'invariance fonctionnelle, les **estimateurs du MV** vérifient aussi la relation $\hat{\beta}_{(Y)} = -\hat{\beta}_{(Z)}$ dans le cas des modèles LOGIT et PROBIT On peut également s'en persuader directement en remarquant que la symétrie de la fonction F permet d'écrire $F(-X_i\beta) = 1 - F(X_i\beta)$, de sorte que la log-vraisemblance vérifie :

$$l = L_I(Y_i, \beta) = \sum_{i=1}^{I} Y_i \log(F(X_i\beta)) + \sum_{i=1}^{I} (1 - Y_i) \log(1 - F(X_i\beta)) = L_I(1 - Y_i, -\beta)$$

Concrètement, vous avez la possibilité, soit de changer le signe de vos coefficients lorsque vous donnez vos résultats, soit d'utiliser l'option `Descending` de la `Proc Logistic`, soit enfin de recodifier au début du programme :

$$Y = \begin{cases} 1 & \text{je possède le bien} \\ 2 & \text{je ne le possède pas} \end{cases}$$

C'est cette seconde solution qui est présentée dans les exemples.

♆ L'interprétation de la constante

Par ailleurs, l'interprétation de la constante est délicate et nécessite de revenir sur les fondements de la méthode. En matière d'analyse des comportements, on fait habituellement l'hypothèse que le choix du consommateur est régi par une variable latente, qui représente la propension qu'il a à réaliser la pratique (section III). C'est sur cette propension que l'on postule le modèle linéaire. De fait, si elle était directement observée, on se trouverait dans le cas usuel de l'analyse de variance sur variable expliquée quantitative. Mais ce que l'on observe pratiquement correspond au fait que la propension dépasse un certain seuil. Le consommateur choisit donc de pratiquer si sa propension est au-delà du seuil s, de ne pas pratiquer sinon :

$$Z > s \Leftrightarrow Y = 1$$

Or,

$$Z = a + Xb + u$$

si l'on isole la constante parmi les variables explicatives. De ce fait,

$$Z > s \Leftrightarrow a - s + Xb + u > 0$$

Le modèle ainsi spécifié n'est donc pas identifiable. Seul l'écart entre la constante et le seuil l'est. Il correspond à la variable INTERCPT dans l'estimation, qui est donc difficilement interprétable en soi. Dans la lecture des résultats, plus la pratique est rare (modalité 2 fréquente), plus le seuil est élevé, et donc plus le coefficient de la variable INTERCPT est négatif. Inversement, plus la pratique est fréquente, plus ce coefficient est positif. Cette difficulté ne pose qu'un problème d'interprétation de la constante. Elle est sans conséquence lorsque l'on cherche à recalculer les probabilités estimées selon les caractéristiques individuelles, et notamment celle qui correspond à la situation de référence.

♆ Dans le cas de données de départ très nombreuses,

il est possible de travailler sur une table croisée au lieu de travailler sur la table complète des observations, en utilisant la syntaxe dite « événements/expériences » (events/trials) de la `Proc Logistic`. En effet, l'exhaustivité des statistiques $(X_i Y_i, N_i)$, (voir la vraisemblance) permet de travailler sur le tableau issu d'une `Proc Summary` de la variable expliquée, triée selon les variables explicatives.

Cela n'a d'intérêt que si toutes les variables explicatives sont discrètes : dans ce cas, on a un nombre fini (et relativement limité) de croisements des variables explicatives, et `_FREQ_` est alors pris comme nombre d'essais (trials) `SUM` est pris comme le nombre d'événements (events). La procédure tourne alors beaucoup plus rapidement que sur la table des observations.

2. Quelques rappels de syntaxe

Pour plus de détails, voir la brochure SAS intitulée SAS/STAT User's Guide volume 2 version 6. La syntaxe est la suivante :

```
Proc Logistic <options 1>;
Model Y = X1X2. . .</options 2>;
```

instructions obligatoires

```
By variables;
Test équation<, équation, équation>;
Output <Out= table sas> <mot clé = nom1 mot clé = nom2> </Alpha= valeur>;
Weight variable;
```

instructions facultatives

Les parties entre < > sont optionnelles.

Parmi les options 1 :

• `Data` = pour préciser la table SAS où sont les données de départ (par défaut le dernier créé)

• des options pour modifier les impressions automatiques

• `Outest` = crée une table SAS qui contient les estimateurs définitifs des paramètres et en option leur covariance estimée. Dans le cas d'un modèle dichotomique, les noms des variables dans cette table sont les mêmes que ceux des variables explicatives de `MODEL` plus le nom INTERCEP pour l'estimateur de la constante.

Parmi les options 2 :

• `Link` = permet de traiter le modèle PROBIT (`Link = Normit`) est celui lié à la loi de Gompertz (`Link = Cloglog`). Par défaut, `Link = Logit`.

• `Noint` ajuste un modèle sans terme constant

• `Selection` = pour sélectionner la méthode de construction du modèle. Par `défaut Selection = None` (l'ajustement se fait sur toutes les variables explicatives indiquées). On peut adopter `Selection = Backward, Selection = Forward, Selection = Stepwise`. D'autres options précisent les impressions désirées quand `Selection` = est précisé (état de départ, niveaux de significativité désirés pour qu'une variable soit retenue...)

• `Ctable` imprime une table de classification (voir plus haut), pour différentes valeurs du seuil définies par SAS. Cette option par défaut peut être modifiée par l'option `Pprob` = .

• Divers diagnostics sur la régression. En particulier, `Iplots` donne des graphiques représentant pour chaque observation la valeur d'un certain nombre de statistiques. Attention quand vous avez beaucoup d'observations !

• `Maxiter` = permet de modifier le nombre d'itérations (cf infra). Le nombre par défaut est 25.

L'instruction TEST :

Cette instruction permet de réaliser des tests de Wald pour toutes les contraintes linéaires, en particulier :

- la nullité d'un coefficient
- la nullité d'un ensemble de coefficients
- l'égalité de deux coefficients
- une ou plusieurs relations linéaires entre coefficients.

Ainsi, après une instruction du style :

```
Proc Logistic Data=TOTO;
Model Y=a1-a5 b2-b6 c1 c2 d1-d4;
```

on peut programmer :

TEST a1=0; (ou TEST a1;)	test de la nullité du coefficient de a1.
TEST a1=0, a2=0,a3=0;	test de nullité conjointe des coefficients de a1, a2 et a3.
TEST a2=a4;	test d'égalité des coefficients de a2 et a4
TEST a1-2*a3=d1-b4, a5=4*b3 ;	test d'un système de relations linéaires entre coefficients

3. Quelques précisions sur les procédures de sélection pas à pas des variables

Procédure FORWARD

Cette procédure entre les variables une à une dans le modèle. On peut partir d'un modèle avec constante seulement (c'est ce qui est fait par défaut) ou spécifier des variables incluses obligatoirement dans le modèle, par les instructions `START` et `INCLUDE`.

SAS procède alors à l'aide de l'algorithme suivant :

⇨ La procédure `LOGISTIC` estime d'abord les paramètres pour les variables présentes dans le modèle.

⇨ La procédure calcule ensuite pour chaque variable non présente dans le modèle, la statistique du « Khi-deux résiduel », c'est-à-dire la statistique du score pour le test :

Ho : modèle comprenant toutes les variables entrées jusqu'à cette étape.

H1 : modèle comprenant toutes les variables entrées jusqu'à cette étape plus la variable examinée.

- Si une de ces statistiques est significative au niveau indiqué en entrée par `SLENTRY` = (par défaut, 0.05), la variable pour laquelle la statistique est la plus grande est entrée dans le modèle. On revient à l'étape d'estimation pour le modèle augmenté.

- Sinon, la procédure est terminée, et le modèle retenu est celui de la dernière étape.

Exemple de mise en oeuvre avec comme variable de départ la seule constante. On notera que la procédure fournit différents tests (rapport de vraisemblance,...), mais que le seul utilisé pour retenir les variables est le tests du score entre la constante et les variables à introduire ((3) et (7) dans l'exemple) :

```
Forward Selection Procedure

Step  0. Intercept entered:

                Residual Chi-Square = 4526.9889 with 29 DF (p=0.0001) (1)
```

(1) Test du score pour

Ho : modèle avec la constante seule

contre H1 : modèle avec toutes les variables explicatives (29 degrés de liberté)

Step 1. Variable PROP entered:

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	33334.145	31323.487	.
SC	33342.422	31340.040	.
-2 LOG L	33332.145	31319.487	2012.658 with 1 DF (p=0.0001) (2)
Score	.	.	1999.200 with 1 DF (p=0.0001) (3)

Residual Chi-Square = 2637.4985 with 28 DF (p=0.0001) (4)

(2) Test du rapport de vraisemblance pour
Ho : modèle avec la constante seule
contre H1 : modèle avec la constante et la variable PROP (1 degré de liberté)

(3) Test du score pour les mêmes hypothèses

(4) Test du score pour
Ho : modèle avec la constante et la variable PROP
contre H1 : modèle avec toutes les variables explicatives.

Si Ho est vraie, la statistique du score doit suivre asymptotiquement un Khi-deux à (29-1=28) degrés de liberté

Step 2. Variable IAAT9 entered:

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	33334.145	30741.466	.
SC	33342.422	30766.295	.
-2 LOG L	33332.145	30735.466	2596.679 with 2 DF (p=0.0001) (5)
Score	.	.	2498.556 with 2 DF (p=0.0001) (6)

Residual Chi-Square = 2085.9259 with 27 DF (p=0.0001) (7)

(5) Test du rapport de vraisemblance pour
Ho : modèle avec la constante seule
contre H1 : modèle avec la constante et les variables PROP et IAAT9 (2 degrés de liberté).
C'est un test de significativité des deux variables prises en même temps.

(6) Test du score pour les mêmes hypothèses

(7) Test du score pour
Ho : modèle avec la constante et les variables PROP et IAAT9
contre H1 : modèle avec toutes les variables explicatives.

Si Ho est vraie, la statistique du score doit suivre asymptotiquement un Khi-deux à (29-2=27) degrés de liberté

ETC. la séquence se poursuit jusqu'à ce qu'on ne trouve plus de variable au seuil de significativité donné dans la procédure (ici, 0.01). SAS indique alors ceci :

NOTE: No (additional) variables met the 0.01 significance level for entry into the model.

Summary of Forward Selection Procedure

Step	Variable Entered	Number In	Score Chi-Square (8)	Pr > Chi-Square
1	PROP	1	1999.2	0.0001
2	IAAT9	2	522.6	0.0001
3	CONF1	3	332.9	0.0001
4	SURPEUP	4	254.0	0.0001
5	CONF3	5	214.3	0.0001
6	CONF2	6	222.0	0.0001
7	AUTRE	7	240.0	0.0001
8	RUC10	8	154.8	0.0001
9	IAAT8	9	149.7	0.0001
10	IAAT7	10	159.0	0.0001
11	RUC9	11	85.7378	0.0001
12	TUR8	12	46.1685	0.0001
13	RUC8	13	53.8453	0.0001
14	RUC7	14	69.2643	0.0001
15	RUC6	15	23.9621	0.0001
16	INDIVID	16	15.8364	0.0001
17	IAAT2	17	13.6342	0.0002
18	IAAT1	18	14.1768	0.0002
19	IAAT3	19	20.6976	0.0001
20	IAAT4	20	9.9836	0.0016
21	IAAT5	21	11.9465	0.0005
22	TUR0	22	7.4535	0.0063

Les statistiques du score indiquées dans le tableau ne correspondent pas (sauf la première) à celles apparues dans les étapes précédentes, car elles ne correspondent pas aux mêmes hypothèses nulles et alternatives. Pour la même raison, les scores ne sont pas systématiquement décroissants dans le tableau.

(8) Test du score pour
Ho : modèle avec la constante et les variables rentrées jusqu'à l'étape précédente
contre H1 : modèle avec la constante et les variables rentrées jusqu'à l'étape courante.

Si Ho était vraie, la statistique du score aurait dû suivre asymptotiquement un Khi-deux à 1 degré de liberté. La PROC LOGISTIC se base sur ces statistiques pour retenir les variables. La valeur critique au seuil rentré ici (1%) est de 6,63. Pour toutes les variables retenues, la statistique du score dépasse cette valeur.

Procédure BACKWARD

Cette procédure part du modèle complet (ou du modèle comprenant les variables spécifiées dans l'instruction START ou INCLUDE) et élimine les variables 1 à 1 du modèle. Par défaut, le modèle de départ est le modèle complet.

SAS procède alors à l'aide de l'algorithme suivant :

⇨ (1) La procédure LOGISTIC estime les paramètres pour les variables encore présentes dans le modèle. On passe en (2).

⇨ (2) - Si toutes les variables sont significatives individuellement (au sens du test de Wald), au niveau indiqué par SLSTAY = (par défaut, 0.05), la procédure s'arrête.
- Si une des variables n'est pas significative individuellement, la moins significative est éliminée du modèle. On passe en (1).

Le test de suppression des variables n'est donc pas le même que pour la procédure FORWARD. De même, il n'y a aucune raison pour que les variables enlevées *in fine* soient exactement celles qui ne sont pas significatives dans le modèle complet, puisque les statistiques de Wald qui entrent en jeu sont celles du modèle de l'étape courante. (voir *infra* le tableau de comparaison des effets des différentes procédures).

Backward Elimination Procedure

Step 0. The following variables were entered:

INTERCPT	TUR0	TUR1	TUR4	TUR8	INDIVID	PROP	AUTRE
IAAT1	IAAT2	IAAT3	IAAT4	IAAT5	IAAT7	IAAT8	IAAT9
RUC1	RUC2	RUC3	RUC4	RUC6	RUC7	RUC8	RUC9
RUC10	SURPEUP	SOUPEUP	CONF1	CONF2	CONF3		

Step 1. Variable SOUPEUP is removed:

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	33334.145	28669.719	.
SC	33342.422	28909.739	.
-2 LOG L	33332.145	28611.719	4720.426 with 28 DF (p=0.0001)
Score	.	.	4526.249 with 28 DF (p=0.0001)

Residual Chi-Square = 0.6026 with 1 DF (p=0.4376) (9)

(9) Test du score pour

Ho : modèle avec la constante et toutes les variables sauf SOUPEUP

contre H1 : modèle avec toutes les variables explicatives .

Si Ho est vraie, la statistique du score doit suivre asymptotiquement un Khi-deux à 1 degré de liberté. Ici, la valeur 0.6026 a 43,76 % de chance d'être dépassée. On accepte donc la validité de Ho au seuil de 1%. (ATTENTION : ce n'est pas ce test qui est utilisé par SAS dans la procédure BACKWARD. Ce test est seulement un moyen de contrôle !).

.....

Step 5. Variable TUR0 is removed:

Residual Chi-Square = 13.7857 with 5 DF (p=0.0170)

NOTE: No (additional) variables met the 0.01 significance level for removal from the model.

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square (1)	Pr > Chi-Square
1	SOUPEUP	28	0.6026	0.4376
2	RUC6	27	0.7696	0.3803
3	TUR1	26	0.9986	0.3176
4	TUR4	25	4.8727	0.0273
5	TUR0	24	6.5954	0.0102

(1) Valeur de la statistique de Wald pour la variable enlevée, dans le modèle de l'étape courante. C'est le critère d'élimination utilisé par SAS.

Procédure STEPWISE

C'est une combinaison des deux procédures précédentes. A chaque étape, SAS regarde s'il peut ajouter une variable (comme dans FORWARD): si c'est le cas, il calcule les paramètres et leurs écarts-types. Si une ou plusieurs variables ne sont pas significatives individuellement, une élimination suivant les principes de la procédure BACKWARD intervient (une ou plusieurs variables sont éliminées). Et ainsi de suite.
Dans cette procédure, deux seuils interviennent : le seuil d'acceptation d'une variable (SLENTRY=) et le seuil d'élimination (SLSTAY=). Par défaut, ces deux seuils sont fixés à 0.05.

Summary of Stepwise Procedure

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	PROP		1	1999.2	.	0.0001
2	IAAT9		2	522.6	.	0.0001
3	CONF1		3	332.9	.	0.0001
4	SURPEUP		4	254.0	.	0.0001
5	CONF3		5	214.3	.	0.0001
6	CONF2		6	222.0	.	0.0001
7	AUTRE		7	240.0	.	0.0001
8	RUC10		8	154.8	.	0.0001
9	IAAT8		9	149.7	.	0.0001
10	IAAT7		10	159.0	.	0.0001
11	RUC9		11	85.7378	.	0.0001
12	TUR8		12	46.1685	.	0.0001
13	RUC8		13	53.8453	.	0.0001
14	RUC7		14	69.2643	.	0.0001
15	RUC6		15	23.9621	.	0.0001
16	INDIVID		16	15.8364	.	0.0001
17	IAAT2		17	13.6342	.	0.0002
18	IAAT1		18	14.1768	.	0.0002
19	IAAT3		19	20.6976	.	0.0001
20	IAAT4		20	9.9836	.	0.0016
21	IAAT5		21	11.9465	.	0.0005
22	TUR0		22	7.4535	.	0.0063

Dans ce cas, il n'y a pas eu de variable enlevée. Les variables retenues sont donc les mêmes (et elles sont rentrées dans le même ordre) que dans la procédure FORWARD.

Comparaison des variables retenues par les options FORWARD et BACKWARD

Le tableau suivant montre pour les mêmes données les variables retenues par les options FORWARD et BACKWARD. En grisé, sont indiqués les variables dont les coefficients ne sont pas significatifs dans le modèle initial.

Modèle complet	BACKWARD	FORWARD
TUR0		X
TUR1		
TUR4		
TUR8	X	X
INDIVID	X	X
PROP	X	X
AUTRE	X	X
IAAT1	X	X
IAAT2	X	X
IAAT3	X	X

IAAT4	X	X
IAAT5	X	X
IAAT7	X	X
IAAT8	X	X
IAAT9	X	X
RUC1	X	
RUC2	X	
RUC3	X	
RUC4	X	
RUC6		X
RUC7	X	X
RUC8	X	X
RUC9	X	X
RUC10	X	X
SURPEUP	X	X
SOUPEUP		
CONF1	X	X
CONF2	X	X
CONF3	X	X

On observe des différences entre les variables retenues à l'aide du modèle complet, des options FORWARD et BACKWARD.

De toute façon, les procédures de sélection automatique ne dispensent pas de réfléchir : par exemple, il faut prendre garde au fait que les variables retenues dépendront crucialement du choix de la situtation de référence.

4. un exemple de sortie interprétée

La variable dépendante est IPOLLU, où :

IPOLLU = 1 si le ménage souffre de la pollution
IPOLLU = 2 s'il n'en souffre pas

Les variables explicatives sont :

DIP2 à DIP5	: Diplôme de la personne de référence du ménage
STR2 à STR5	: catégorie de commune
STL1,STL3	: statut d'occupation du logement
AGE1 à AGE3, AGE5, AGE6	: âge de la personne de référence
REV1 à REV4, REV6	: tranche de revenu du ménage

Auxquelles s'ajoute la constante (intercept) représenté par INTERCPT. La situation de référence est donc définie par la nullité des coefficients des variables DIP1, STR1, STL2, AGE4, REV5.

Nous n'avons pas demandé dans la sortie quelques statistiques descriptives élémentaires portant sur les variables explicatives, peu intéressantes quand il s'agit de variables dichotomiques. Pour cela, il aurait fallu utiliser l'option `Simple`.

La moyenne -mean- combinée avec le nombre d'observations du fichier dans un modèle non pondéré permet toutefois de retrouver les effectifs de chaque modalité des variables explicatives. Il paraît plus simple de faire toujours précéder le modèle d'une `Proc Freq` sur les modalités des variables explicatives. Bien qu'il n'y ait pas de limite inférieure à respecter sur ces effectifs, il conviendra d'être prudent quant à l'interprétation des coefficients estimés sur des strates d'effectif réduit (de l'ordre de moins de 20).

Le programme était donc :

```
Proc Logistic Data = codif ;
Model IPOLLU    = DIP2-DIP5
                  STR2-STR5
                  STL1 STL3
                  AGE1-AGE3 AGE5 AGE6
                  REV1-REV4 REV6 / Ctable ;
```

The LOGISTIC Procedure *Exemple 1a :Logit*

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution ⇒*Variable Dépendante*
Response Levels: 2
Number of Observations: 7332
Link Function: Logit ⇒*Modèle Logit*

Response Profile

Ordered Value	IPOLLU	Count
1	1	1161
2	2	6171

Criteria for Assessing Model Fit ⇒*Critères permettant de juger de l'ajustement du modèle*

Criterion	Intercept Only ⇒*1*	Intercept and Covariates ⇒*2*	Chi-Square for Covariates ⇒*2 Modèle avec constante et variables X*
AIC	6408.974	6291.197	.
SC	6415.874	6436.097	.
-2 LOG L	6406.974	6249.197	157.777 with 20 DF (p=0.0001) ⇒*3*
Score	.	.	150.077 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio ⇒*4*
INTERCPT	1	-2.3262	0.1453	256.3715	0.0001	.	0.098
DIP2	1	0.0884	0.1070	0.6830	0.4085	0.019039	1.092
DIP3	1	0.0236	0.1095	0.0463	0.8297	0.005061	1.024
DIP4	1	-0.1521	0.1315	1.3386	0.2473	-0.026394	0.859
DIP5	1	-0.0156	0.1062	0.0215	0.8835	-0.003918	0.985
STR2	1	0.5374	0.1162	21.3720	0.0001	0.110465	1.712
STR3	1	0.5918	0.1216	23.6987	0.0001	0.112738	1.807
STR4	1	0.9945	0.1011	96.6763	0.0001	0.249008	2.703
STR5	1	0.9641	0.1181	66.6684	0.0001	0.186275	2.622
STL1	1	0.1612	0.0897	3.2334	0.0721	0.043923	1.175
STL3	1	0.1680	0.1021	2.7058	0.1000	0.042613	1.183
AGE1	1	-0.0688	0.1260	0.2987	0.5847	-0.012528	0.933
AGE2	1	0.0193	0.1073	0.0325	0.8570	0.004227	1.020
AGE3	1	0.0139	0.1016	0.0188	0.8910	0.003132	1.014
AGE5	1	-0.1417	0.1119	1.6037	0.2054	-0.027505	0.868
AGE6	1	-0.3287	0.1391	5.5810	0.0182	-0.052184	0.720
REV1	1	-0.1619	0.1383	1.3694	0.2419	-0.024539	0.851
REV2	1	-0.1278	0.1105	1.3362	0.2477	-0.025827	0.880
REV3	1	-0.1462	0.1196	1.4922	0.2219	-0.025890	0.864
REV4	1	-0.0569	0.1048	0.2952	0.5869	-0.011459	0.945
REV6	1	-0.0465	0.0946	0.2413	0.6233	-0.010947	0.955

Association of Predicted Probabilities and Observed Responses ⇒*5*

Concordant = 60.7%	Somers' D = 0.228
Discordant = 37.9%	Gamma = 0.231
Tied = 1.4%	Tau-a = 0.061
(7164531 pairs)	c = 0.614

The LOGISTIC Procedure

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.060	1161	0	6171	0	15.8	100.0	0.0	84.2	.
0.080	1111	433	5738	50	21.1	95.7	7.0	83.8	10.4
0.100	1011	1533	4638	150	34.7	87.1	24.8	82.1	8.9
0.120	973	1857	4314	188	38.6	83.8	30.1	81.6	9.2
0.140	885	2378	3793	276	44.5	76.2	38.5	81.1	10.4
0.160	728	3317	2854	433	55.2	62.7	53.8	79.7	11.5
0.180	607	3843	2328	554	60.7	52.3	62.3	79.3	12.6
0.200	453	4425	1746	708	66.5	39.0	71.7	79.4	13.8
0.220	197	5377	794	964	76.0	17.0	87.1	80.1	15.2
0.240	31	6048	123	1130	82.9	2.7	98.0	79.9	15.7
0.260	0	6171	0	1161	84.2	0.0	100.0	.	15.8

⇒*Exemple avec un seuil de 0.140:*

885 observations ont une valeur 1 pour ipollu réelle et prédite
2378 ont la valeur réelle 2 et la valeur prédite 2
3793 ont la valeur réelle 2 et la valeur prédite 1
276 ont la valeur réelle 1 et la valeur prédite 2

Correct = (885+2378)/(885+2378+3793+276)
Sensisivity = 885/(885+276)
Spécificity = 2378/(2378+3793)
False POS = 3793/(3793+885)
False NEG = 276/(276+2378)

⇒3: Valeur de $-2(\log(L_1)-\log(L_2))$

La probabilité que le χ^2 à 20 degrés de liberté dépasse cette valeur est de p=0.0001.

L'hypothèse nulle (les variables explicatives autres que la constante n'expliquent pas les disparités) est donc rejetée.

⇒4: `Parameter Estimate`: estimateur du paramètre $\hat{\beta}_j$

`Standard Error`: écart-type du paramètre $\hat{\sigma}_j$

`Wald Chi-Square`: statistique de Wald ; si $\frac{\hat{\beta}_j}{\hat{\sigma}_j} > 2^2 = 4$, le paramètre est non nul.

`Pr > Chi-Square`: le coefficient est significativement non nul si cette probabilité est inférieure à 0.05

`Standardized estimate`: estimateur standardisé $\frac{\hat{\beta}_j}{r_j}$,

où r_j est le rapport entre l'écart-type de la fonction de répartition de la loi logistique (normale si `Link=Normit`) et l'écart-type de la jème variable explicative dans l'échantillon. L'estimateur normalisé permet notamment de comparer les estimateurs des modèles logit et probit.

`Odds Ratio`: correspond à $\exp(\hat{\beta}_j)$ dans le cas du modèle dichotomique

⇒5: proportion des paires concordantes et discordantes (voir le texte sur les tests et autres indicateurs de validité du modèle).

5. Le fichier en sortie

Pour obtenir un fichier (une table SAS) en sortie, il faut faire appel à une instruction facultative, l'instruction Output.

On écrira alors :

```
Proc Logistic ;
Model Variable dépendante = variables explicatives ;
Output Out = nom de la table SAS en sortie
             <mot-clé = nom ... mot-clé = nom> ;
```

Le fichier en sortie est une nouvelle table SAS qui contient toutes les variables de la table en entrée. En option, l'instruction Output crée l'estimateur $X\hat{\beta}$ de la partie linéaire du modèle, son écart-type estimé, la probabilité estimée pour chaque individu d'avoir la modalité la plus faible $Y = 1$, l'intervalle de confiance pour cette probabilité, et des statistiques d'aide au diagnostic sur la régression.

Pour obtenir en sortie, par exemple, la probabilité estimée (que l'individu ait pour Y la valeur la plus faible) on emploiera le mot-clé `Predicted`, ou `P`. Si on veut lui donner le nom EQUIP, on écrira :

```
Proc Logistic ;
Model ... ;
Output Out = ... P = EQUIP ;
```

Toutefois pour obtenir la probabilité $\hat{p}$ estimée, il est plus simple d'utiliser l'option `Outest` de l'instruction `Proc Logistic` et de calculer $\hat{p}$ pour les modalités qui nous intéressent.

Exemple : si on a : IPOLLU = 1 si le ménage souffre de la pollution
IPOLLU = 2 s'il n'en souffre pas

```
Proc Logistic Data = codif Outest = TAB;
Model IPOLLU   = DIP2-DIP5
                 STR2-STR5
                 STL1 STL3
                 AGE1-AGE3 AGE5 AGE6
                 REV1-REV4 REV6 / Ctable;
Output Out = POLLU1 P = PHAT;
```

On veut obtenir la probabilité estimée de souffrir de la pollution des ménages pour lesquels les variables DIP2, STR5, STL1, AGE2 et REV3 valent 1.

```
1. Data A ; Set TAB ;
X1 = -(INTERCEP+DIP2+STR5+STL1+AGE2+REV3) ;
PHAT1 = 1/ (1 + Exp (X1)) ;
Proc Print Data = A ; Var PHAT1 ;

2. Data B ; Set POLLU1 ;
If DIP2 = 1 & STR5 = 1 & STL1 = 1 & AGE2 = 1 & REV3 = 1;
Proc Print Data = B  (Obs = 1) ; Var PHAT ;
```

Les valeurs qu'on obtient par PHAT et PHAT1 sont égales et représentent $\hat{p}$. Une mise en garde toutefois : dans `POLLU1`, comme dans `CODIF`, les variables DIP2 à REV6 sont les variables explicatives (0 ou 1) du modèle ; dans TAB, il s'agit des coefficients estimés. La solution 2 n'est en outre pas applicable lorsque la situation définie par la conjonction des conditions n'existe pas dans l'échantillon, ce qui peut parfois se produire. Même si les calculs sont analytiquement corrects dans la solution 1, on peut s'interroger sur la pertinence d'un cas de figure aussi rare (« dromadaires sur la banquise », cf infra.). L'hypothèse d'additivité atteint là ses limites, comme le statisticien utilisateur.

Pour la question de l'utilisation des pondérations lors du calcul de probabilités estimées, voir « Pondérer ou ne pas pondérer », dans la partie IX.

6. Modèle LOGIT, modèle PROBIT

Le modèle PROBIT est traité dans SAS, on l'a vu, en ajoutant l'option `Link = Normit` à l'instruction `Model`.

Les résultats obtenus ne sont pas directement comparables. Il faut comparer les estimateurs standardisés (*standardized estimates*) qui tiennent compte de la différence de variance entre les deux distributions.

Le programme était ici :

```
Proc Logistic Data = codif ;
Model IPOLLU = les mêmes variables
                / Ctable Link = Normit ;
```

Le lecteur se convaincra aisément que même si certains coefficients standardisés diffèrent légèrement, les conclusions qu'ils permettent de tirer sont identiques.

The SAS System

The LOGISTIC Procedure ⇒*Exemple 1b Probit*

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution
Response Levels: 2
Number of Observations: 7332
Link Function: Normit ⇒*Modèle PROBIT*

Response Profile

Ordered Value	IPOLLU	Count
1	1	1161
2	2	6171

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	6408.974	6291.565	.
SC	6415.874	6436.465	.
-2 LOG L	6406.974	6249.565	157.409 with 20 DF (p=0.0001)
Score	.	.	150.077 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	-1.3408	0.0779	296.5953	0.0001	.
DIP2	1	0.0429	0.0586	0.5364	0.4639	0.016762
DIP3	1	0.0171	0.0602	0.0804	0.7767	0.006650
DIP4	1	-0.0854	0.0720	1.4095	0.2351	-0.026886
DIP5	1	-0.00832	0.0587	0.0201	0.8872	-0.003801
STR2	1	0.2806	0.0609	21.2386	0.0001	0.104600
STR3	1	0.3091	0.0642	23.1760	0.0001	0.106784
STR4	1	0.5367	0.0534	101.0471	0.0001	0.243723
STR5	1	0.5189	0.0637	66.3664	0.0001	0.181842
STL1	1	0.0857	0.0495	3.0024	0.0831	0.042362
STL3	1	0.0908	0.0560	2.6246	0.1052	0.041762
AGE1	1	-0.0434	0.0702	0.3830	0.5360	-0.014340
AGE2	1	0.00657	0.0596	0.0122	0.9122	0.002605
AGE3	1	0.000818	0.0564	0.0002	0.9884	0.000334
AGE5	1	-0.0770	0.0617	1.5538	0.2126	-0.027087
AGE6	1	-0.1770	0.0749	5.5778	0.0182	-0.050968
REV1	1	-0.0917	0.0759	1.4568	0.2274	-0.025204
REV2	1	-0.0682	0.0606	1.2643	0.2608	-0.025001
REV3	1	-0.0803	0.0656	1.4977	0.2210	-0.025811
REV4	1	-0.0328	0.0581	0.3182	0.5727	-0.011958
REV6	1	-0.0260	0.0528	0.2424	0.6225	-0.011111

Association of Predicted Probabilities and Observed Responses

Concordant = 60.7%	Somers' D = 0.227
Discordant = 38.0%	Gamma = 0.230
Tied = 1.4%	Tau-a = 0.060
(7164531 pairs)	c = 0.613

The SAS System

The LOGISTIC Procedure

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.040	1161	0	6171	0	15.8	100.0	0.0	84.2	.
0.060	1161	1	6170	0	15.8	100.0	0.0	84.2	0.0
0.080	1118	457	5714	43	21.5	96.3	7.4	83.6	8.6
0.100	1015	1514	4657	146	34.5	87.4	24.5	82.1	8.8
0.120	975	1866	4305	186	38.7	84.0	30.2	81.5	9.1
0.140	888	2368	3803	273	44.4	76.5	38.4	81.1	10.3
0.160	737	3318	2853	424	55.3	63.5	53.8	79.5	11.3
0.180	620	3828	2343	541	60.7	53.4	62.0	79.1	12.4
0.200	470	4369	1802	691	66.0	40.5	70.8	79.3	13.7
0.220	195	5384	787	966	76.1	16.8	87.2	80.1	15.2
0.240	26	6076	95	1135	83.2	2.2	98.5	78.5	15.7
0.260	0	6171	0	1161	84.2	0.0	100.0	.	15.8

IX Mise en oeuvre du modèle LOGIT

Cette partie vise à présenter une sorte de « check list » recensant les questions successives que doit se poser le statisticien qui, disposant d'un fichier de données individuelles, désire étudier un phénomène à l'aide du modèle Logit. En fait la plupart des remarques présentées s'appliquent tout autant aux modèles d'analyse de variance classiques qu'au modèle Logit et forment un mode d'emploi assez général pour qui veut se lancer dans la voie de la séparation des effets, de l'analyse toutes choses égales par ailleurs.

Contrairement au reste de la note, beaucoup plus « objectif », cette partie est nourrie des expériences personnelles de quelques statisticiens : elle reflète donc des prises de position qui peuvent ne pas être partagées par tous ... libre à chacun de s'en démarquer si le sujet étudié le nécessite.

De plus, on peut distinguer deux types d'usage d'un tel modèle : on peut se contenter d'expliquer la variance du phénomène, afin par exemple de le prévoir au mieux ; on peut aussi être plus ambitieux et souhaiter utiliser le modèle pour dégager des processus explicatifs, des relations causales. Certaines difficultés évoquées ci-dessous (en particulier colinéarité, non exogénéité) ne sont réellement gênantes que dans la seconde optique (celle que nous privilégions ici) ; dans la première optique elles peuvent être quasiment négligées.

Un peu de vocabulaire préliminaire : dans la présentation théorique toutes les variables X_j introduites sont placées sur le même plan : le fait d'avoir un revenu situé dans la deuxième tranche et le fait d'avoir un revenu situé dans la cinquième tranche sont traités comme deux variables différentes, au même titre que le fait d'avoir 2 enfants et le fait d'habiter une commune rurale.

Or bien évidemment toutes les variables obtenues par discrétisation d'une même variable continue ou les variables correspondant aux diverses modalités d'une variable qualitative ont entre elles des liens organiques étroits tout à fait spécifiques. Il est utile pour la clarté de l'exposé de tenir compte de ce phénomène au niveau du vocabulaire utilisé : on parlera donc de **dimensions explicatives**, représentées chacune par diverses **variables explicatives** : ainsi le milieu socioprofessionnel sera une dimension explicative. Son introduction dans le modèle se traduira par l'introduction de plusieurs variables explicatives issues directement ou non des variables présentes dans le fichier : « être agriculteur », « être ouvrier », « être cadre supérieur » etc ...

Dans le cas des dimensions explicatives correspondant à une représentation continue, comme le revenu ou l'âge, plusieurs modélisations sont envisageables. Certaines utilisent une seule variable explicative, d'autres plusieurs (cf infra).

1. La spécification du modèle

Assurément la réflexion préalable autour de la spécification du modèle est le point le plus important pour garantir la qualité des résultats : les choix doivent être raisonnés, en faisant par exemple référence aux analyses sociologiques ou économiques disponibles sur le sujet.

a. retenir ou non une dimension explicative.

Quand on introduit une dimension explicative, on doit être capable de décrire les mécanismes par lesquels elle est susceptible d'agir sur le phénomène étudié, voire de prévoir le signe des coefficients. Une démarche purement heuristique du type « j'introduis dans le modèle comme dimensions explicatives tout ce dont je dispose dans mon fichier et je laisse à une procédure automatique du type BACKWARD par exemple, le soin de choisir » est à proscrire. la démarche doit rester une démarche de vérification d'hypothèses bien spécifiées.

Pour pouvoir être introduite dans le modèle, une dimension explicative doit présenter un caractère d'exogénéité par rapport au phénomène étudié. Parfois -mais rarement- cela va de soi : l'âge de l'individu, son sexe peuvent sans difficulté être considérés comme exogènes pour une étude de comportement !

Le revenu, la catégorie socioprofessionnelle, le diplôme, l'activité, le type d'habitat posent davantage de problèmes. On peut certes supposer qu'il s'agit bien de variables exogènes : à court terme elle s'imposent à l'individu et ne sauraient être modifiées. Pour l'étude de comportements quotidiens l'hypothèse semble raisonnable, même si on peut faire quelques objections : est-il, par exemple, licite de supposer que l'activité professionnelle de la femme est exogène quand on étudie des pratiques comme la couture et le tricot. Le fait de ne pas travailler à l'extérieur et de faire du tricot sont peut-être deux manifestations conjointes d'une même variable latente, que l'on peut qualifier pour faire bref de « goût pour la vie au foyer ». C'est peut-être pour pouvoir être en mesure de faire du tricot tous le jours que la femme étudiée a choisi de ne pas avoir d'activité professionnelle. Dans ce cas, la procédure correcte consisterait à recourir à des **équations simultanées** : deux variables expliquées (activité professionnelle, tricot) et des variables explicatives ne contenant pas l'activité professionnelle. Toutefois le traitement économétrique des équations simultanées avec variables qualitatives est souvent difficile à mettre en oeuvre. Cependant, quelques techniques sont disponibles, qui réalisent l'analogue des tests d'Hausmann sur les modèles quantitatifs (voir infra).

Concrètement, il faut souligner que la plupart du temps il n'est guère possible d'affirmer ou de vérifier qu'une dimension explicative est ou n'est pas exogène. Une raison est souvent le manque de variables utilisables pour mettre en place ces tests. Ainsi, en coupe instantanée. on ne dispose généralement pas d'instrument, c'est à dire de variables elles-mêmes exogènes, non utilisées dans la régression, mais raisonnablement corrélées avec les variables suspectées d'endogénéité. Sans pouvoir le vérifier, le statisticien est alors conduit à **postuler** l'exogénéité ou l'endogénéité. Introduire la dimension revient en fait à admettre une exogénéité de façon implicite. Il faut en être conscient, et de préférence, discuter explicitement le problème et ne pas éluder la difficulté : spécifier un modèle, c'est toujours émettre des hypothèses. Ce qui est condamnable, c'est de ne pas donner au lecteur le moyen de les percevoir et de les discuter.

Dans le cas de l'étude de pratiques dont l'horizon temporel est long (achat d'un logement par exemple), supposer que le revenu ou la profession sont exogènes devient cependant très hardi : on peut raisonnablement penser que le comportement de l'agent est conditionné par des variables cachées, comme avoir un but dans la vie (dont l'achat de logement fait partie) et qu'il choisit son intensité de travail ou sa profession de façon à pouvoir réaliser ce but. Il n'y aurait donc pas alors existence d'un revenu exogène venant contraindre les choix sans « effet de retour ».

b. représentation d'une dimension explicative retenue.

Lorsqu'une dimension explicative a été retenue (elle est susceptible d'avoir une influence sur le phénomène étudié et peut raisonnablement être considérée comme exogène), il reste à définir comment la « représenter ».

• Quelles variables pour une dimension ?

↳ Dimension qualitative

Dans le cas d'une dimension explicative de nature qualitative, aucun problème particulier ne se pose. La variable disponible dans le fichier est en général un code à plusieurs modalités. On introduit autant de variables dichotomiques (« dummies ») qu'il y a de modalités. On procède de même avec des variables quantitatives disponibles en tranches. On verra ultérieurement (cf problèmes de non convergence) que l'on peut être amené à effectuer des regroupements de modalités.

↳ Dimension quantitative

Dans le cas d'une dimension explicative de nature continue, différentes possibilités se présentent. Il convient alors de s'interroger sur la façon dont elle intervient.

La première solution revient à faire l'hypothèse de la **linéarité** de l'influence de la dimension explicative sur le phénomène latent étudié, d'un extrême à l'autre de son domaine de variation. La variable peut aussi être introduite sous forme de logarithme pour estimer une sorte d'élasticité. Le modèle s'écrit alors:

$$Z = \alpha + \beta R + \varepsilon$$

où R est la variable Revenu introduite sous sa forme continue.

Un variante consiste à prendre en compte une dépendance quadratique de façon à étudier des sortes de rendement d'échelle. Cette spécification est fréquemment utilisée dans les études académiques et les publications scientifiques. Son avantage est de limiter le nombre de coefficients estimés ; son inconvénient est de contraindre assez fortement la dépendance a priori.

$$Z = \alpha + \beta R + \gamma R^2 + \varepsilon$$

La deuxième solution consiste à se ramener au cas d'une variable qualitative en fabriquant une variable en tranches, et à introduire dans l'estimation autant de variables logiques que de tranches (moins une pour assurer l'identification). La dépendance est moins contrainte que précédemment, mais elle est approximée au moyen d'une fonction en escalier, ce qui peut être gênant si l'on a en tête une représentation continue. Le modèle s'écrit alors:

$$Z = \alpha + \lambda_1 R_1 + \lambda_2 R_2 + \ldots + \lambda_k R_k + \varepsilon$$

où $R_1, \ldots, R_k$ désignent les k variables dichotomiques issues de la discrétisation de R (i. e les k tranches de revenu).

Dans le premier cas, un écart infinitésimale de revenu dR entre deux ménages aura, quel que soit le niveau de départ, un effet $dZ = \beta dR$. Dans le second cas, un écart infinitésimal de revenu (pour deux ménages d'une même tranche) a un effet nul. Un écart de revenu correspondant à un changement de tranche entraîne une variation de Z égale à la différence entre les coefficients. **Si l'on exploite une enquête en coupe instantanée, on se gardera néanmoins d'interpréter un constat établi en termes de statique comparative comme une projection dynamique. A moins de postuler explicitement que les préférences des ménages sont rigoureusement identiques, rien ne permet d'assurer qu'un ménage donné confronté à une variation de revenu se comportera conformément à l'estimation en coupe.**

En général, quand on exploite des enquêtes, on dispose d'échantillons de taille suffisante, et il semble préférable de ne pas postuler d'emblée l'existence d'une dépendance linéaire. Il est indispensable de tester l'hypothèse quadratique quitte à la rejeter si le coefficient du terme quadratique n'est pas significatif, ou si l'introduction de la variable au carré rend le modèle instable. La discrétisation est une solution alternative intéressante puisqu'aucune forme fonctionnelle n'est supposée a priori. Cependant, elle ne permet pas de tester la linéarité de la dépendance puisque l'on est face à une fonction linéaire dans le premier cas et à une fonction en escalier dans le second. C'est pourquoi on préconise une troisième solution, plus complexe mais plus satisfaisante, qui consiste à approximer la dépendance au moyen d'une fonction linéaire par morceaux. La construction des variables explicatives est plus complexe, mais la modélisation est plus souple et sans doute mieux adaptée à la représentation d'un phénomène continu. Elle est en outre plus aisément interprétable et permet de tester la sous-hypothèse linéaire.

Le problème se présente de façon assez analogue à la discrétisation d'une variable en tranches. En effet, on définit à partir de la variable de revenu un certain nombre d'intervalles ($[R_0,R_1],[R_1,R_2],\ldots,[R_k,R_{k+1}],\ldots[R_{K-1},R_K]$) avec le cas échéant $R_0=-\infty$ ou $R_K=+\infty$. La différence est que sur ces intervalles la fonction qui relie la variable latente au revenu est continue et linéaire par morceaux au lieu d'être constante par morceaux :

$$R\in[R_k,R_{k+1}] \qquad Z=\alpha_k+\beta_k R$$

où α_k est défini par continuité entre deux intervalles.

On a alors:

$$\begin{array}{lllllll}
R\in[R_0,R_1] & Z=\alpha & +\beta_0(R-R_1) & & & & \\
R\in[R_1,R_2] & Z=\alpha & & +\beta_1(R-R_1) & & & \\
R\in[R_2,R_3] & Z=\alpha & & +\beta_1(R_2-R_1) & +\beta_2(R-R_2) & & \\
R\in[R_k,R_{k+1}] & Z=\alpha & & +\beta_1(R_2-R_1) & +\beta_2(R_3-R_2) & +\ldots & +\beta_k(R-R_k)
\end{array}$$

Ceci conduit à ajuster un modèle de la forme:

$$Z=\alpha+\beta_0V_0+\beta_1V_1+\ldots+\beta_{K-1}V_{K-1}+\varepsilon$$

qui se présente de façon analogue au modèle issu de la discrétisation d'une variable en tranche. Au lieu d'être des variables dichotomiques, les fonctions V sont définies de la façon suivante:

$$\begin{aligned}
V_0&=(R-R_1)*(R<R_1)\\
V_1&=(R-R_1)*(R_1\le R<R_2)+(R_2-R_1)*(R\ge R_2)\\
V_2&=(R-R_2)*(R_2\le R<R_3)+(R_3-R_2)*(R\ge R_3)\\
V_k&=(R-R_k)*(R_k\le R<R_{k+1})+(R_{k+1}-R_k)*(R\ge R_{k+1})\\
V_{K-1}&=(R-R_{K-1})*(R\ge R_{K-1})
\end{aligned}$$

Dans ces expressions en langage SAS, les parenthèses ont une double signification. Elles assurent le rôle habituel de factorisation pour les produits. Elles permettent également de fabriquer des variables indicatrices (voir supra).

Dans l'exemple, la situation de référence correspond à un revenu égal à R_1. Un changement d'origine (par exemple R_2) modifierait l'écriture de la façon suivante :

$$\begin{aligned}
V_0&=(R-R_1)*(R<R_1)\\
V_1&=(R-R_2)*(R_1\le R<R_2)+(R_1-R_2)*(R\le R_1)\\
V_2&=(R-R_2)*(R_2\le R<R_3)+(R_3-R_2)*(R\ge R_3)\\
V_k&=(R-R_k)*(R_k\le R<R_{k+1})+(R_{k+1}-R_k)*(R\ge R_{k+1})\\
V_{K-1}&=(R-R_{K-1})*(R\ge R_{K-1})
\end{aligned}$$

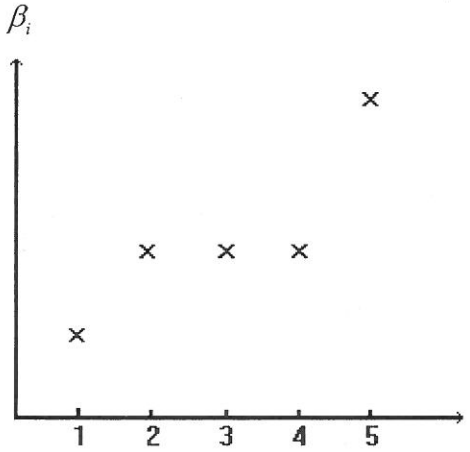
Les deux situations ne diffèrent que par une translation de la constante. Dans les deux cas, l'hypothèse de linéarité ($Z=\alpha+\beta R$) correspond au cas où les coefficients $\beta_0,\beta_1,\ldots,\beta_{K-1}$ ne sont pas significativement différents les uns des autres. Cette hypothèse peut être testée au moyen de la statistique de Wald disponible dans l'instruction `TEST`.

• La situation de référence : à quoi sert-elle ? Comment la choisir ?

Que la dimension explicative soit qualitative (à k modalités), ou continue discrétisée (avec k tranches), se pose le problème suivant : les k variables introduites pour la représenter ne sont pas indépendantes, puisque leur somme vaut 1 quel que soit l'individu i. En effet chaque individu a une CS et une seule, a un revenu qui est dans une tranche et dans une seule. On ne saurait donc estimer les k coefficients, pas plus qu'on ne saurait projeter un vecteur sur deux vecteurs colinéaires.

Le remède consiste à éliminer une des variables. Cette variable non introduite dans le modèle a donc un coefficient égal à 0 par convention et on considère qu'elle représente une **situation de référence**, par rapport à laquelle on mesure des déviations, des différences. Mathématiquement, le choix de cette situation de référence n'a généralement que peu d'importance. Un changement a pour seuls effets une translation des coefficients et une légère modification des écarts-types mesurant la significativité des estimations. Cette dernière n'est gênante que si l'effectif d'une modalité entrant dans la définition de la situation de référence est très faible (quelques unités), ce qui doit être évité. Les effets de la translation des coefficients sont simples : les coefficients changent mais le profil qu'ils dessinent est inchangé ; en particulier l'écart entre le coefficient le plus faible et le plus fort est invariant. En revanche, le nombre de coefficients significativement positifs, négatifs ou de coefficients nuls peut changer, ce qui indique bien que l'on ne peut juger du caractère significatif d'une dimension explicative par le nombre de coefficients non nuls qui apparaissent.

Le graphique suivant permet de visualiser la situation :



Si on choisit la modalité 2 ou (3-4) comme référence, on aura:

$$\begin{cases} \beta_1 = -\beta \\ \beta_2 = \beta_3 = \beta_4 = 0 \\ \beta_5 = +\beta^* \end{cases}$$

Si on choisit la modalité 1, on aura:

$$\begin{cases} \beta_1 = 0 \\ \beta_2 = \beta_3 = \beta_4 = \beta \\ \beta_5 = \beta + \beta^* \end{cases}$$

Si on choisit la modalité 5, on aura:

$$\begin{cases} \beta_1 = -\beta - \beta^* \\ \beta_2 = \beta_3 = \beta_4 = -\beta^* \\ \beta_5 = 0 \end{cases}$$

Dans tous les cas, l'écart maximum vaut $\beta + \beta^*$ mais le nombre de coefficients non nuls varie de 2 (1 positif, 1 négatif) à 4 (tous positifs ou tous négatifs) !

Comme aucun critère mathématique ne vient dicter le choix de la situation de référence, on se laissera guider par des impératifs « esthétiques » : il est plus simple de choisir comme situation de référence une situation courante. Chaque lecteur ainsi acceptera comme naturel le choix qui lui est proposé et le commentaire sera facilité par le fait que l'on opposera des « minorités » bien caractérisées à la population plus « standard ». le précepte conduira souvent à prendre comme référence la situation « modale » (modalité rassemblant le plus d'effectifs) mais il ne s'agit en aucun cas d'une obligation. Rappelons toutefois qu'il est dangereux de choisir une modalité de référence ayant des effectifs trop faibles. Outre la perte de précision déjà mentionnée, cela peut entraîner un défaut de convergence. Ceci étant, une référence ainsi choisie pour chacun des critères peut conduire à une intersection des situations modales très minoritaire dans l'échantillon. Dans les commentaires, il faudra prendre garde à ne pas tomber dans cette « illusion du français moyen »...

c. introduction simultanée de plusieurs dimensions explicatives : problèmes spécifiques à éviter.

Le cas où le modèle se réduit à une seule dimension explicative est très rare. La plupart du temps, la réflexion théorique conduit le statisticien à introduire de très nombreuses dimensions explicatives.

Deux difficultés supplémentaires surgissent, liées aux problèmes de colinéarité et au défaut d'additivité.

• Les problèmes de colinéarité :

Les différentes variables introduites pour représenter une même dimension ne sont pas les seules à être corrélées : en règle générale deux variables quelconques ne sont pas strictement indépendantes.

Mais cela ne pose pas forcément de problèmes au niveau de l'estimation des coefficients.

S'il y a corrélation parfaite (une variable est combinaison linéaire de plusieurs autres) l'identification est impossible, la matrice des variances-covariances n'étant pas inversible. Ceci se présente rarement, et seulement dans des cas où il y a une dépendance logique mécanique entre les variables. On peut citer parmi les cas qui déroutent le plus le novice :

- la corrélation entre le département et un type de commune d'habitat isolant la ville de Paris au sein de l'agglomération parisienne : le département 75 et la ville de Paris intra muros coïncident en effet exactement. Même si elle n'est pas en toute rigueur vérifiée, l'identité entre la région de programme Ile de France et l'ensemble de l'agglomération parisienne peut se produire sur certains échantillons et poser le même type de problèmes.

- la corrélation entre âge et CS, lorsqu'une modalité de ce code isole les retraités : il y a une corrélation mécanique entre retraités et individus très âgés, qui souffre certes des exceptions mais qui peut être parfaite sur des échantillons de taille réduite.

- cas où une dimension explicative est filtrée par une autre dimension, elle aussi introduite dans le modèle. Par exemple, considérons le cas où l'on introduit dans une explication des fortes durées de travail, le type de ménage (pas de couple, couple seul, couple avec autres personnes) et l'existence d'un conjoint exerçant une profession d'indépendant (sans objet ; pas de conjoint ; conjoint salarié ; conjoint indépendant). Les variables « sans objet » et « pas de couple » sont mécaniquement parfaitement corrélées.

Ces exemples paraissent triviaux : ceci traduit bien le fait que les corrélations mécaniques peuvent **toujours** être repérées avec un minimum de réflexion préalable. D'ailleurs, la `Proc Logistic` détecte ce type de corrélation et propose une estimation contrainte.

S'il y a corrélation forte mais non parfaite, non mécanique entre deux variables, il n'y aura pas de problème en général, les échantillons étudiés étant nombreux. Ainsi revenu et diplôme présentent un certain degré de liaison, revenu et diplôme variant plutôt dans le même sens, mais on pourra généralement séparer leurs effets, car il est peu probable qu'il n'y ait pas dans l'échantillon un riche non diplômé ou un diplômé pauvre, ce qui permet l'identification. De même, on pourra généralement séparer l'effet propre de la PCS, des effets de revenu ou de diplôme, malgré les corrélations entre ces variables. Si la liaison est très forte, il est toutefois possible que l'estimation globale du modèle puisse se faire, mais que les coefficients d'une (voire des deux) variables corrélées ne soient pas significativement différents de 0 à cause de ce problème de colinéarité.

D'ailleurs même si l'estimation réussit, la colinéarité pose des problèmes au niveau de l'interprétation des résultats : on ne peut plus parler d'une modification de la situation de référence sur une seule dimension explicative ; il faut gérer des modifications conjointes des dimensions corrélées. Supposons ainsi que l'on ait une variable « être retraité » et une variable « avoir plus de 80 ans » et que ces deux variables ne soient pas suffisamment corrélées pour qu'on ne puisse isoler les deux coefficients (il y a un centenaire actif et un militaire de 45 ans à la retraite), on risquera des erreurs d'interprétation si on oublie qu'en règle générale « avoir plus de 80 ans » entraîne le fait d'être à la retraite et que l'on commente la dimension « âge » indépendamment du reste.

La partie de l'effet « âge élevé », qui est retracée au travers de la variable « être à la retraite» sera omise, et l'effet de l'âge globalement sous-estimé. on ne saurait donc trop recommander d'éliminer au moment de la conception du modèle les colinéarités mécaniques ou quasi mécaniques.

Les solutions à mettre en oeuvre sont simples à définir.

S'il y a redondance pure et simple, on supprime une des deux variables. Dans les exemples précités, ce n'est pas le cas. Dans les 2 premiers cas, le remède consistera à «diluer la corrélation», soit en n'isolant pas la ville de Paris du reste de son agglomération ou en reclassant les inactifs dans leur ancienne profession, soit en fusionnant la variable «Ile de France» et la variable «Bassin Parisien». Dans le dernier cas, on choisira de faire des régressions différentes sur chaque sous-population (une régression pour les temps de travail des couples, une autre pour les non couples, la variable «avoir un conjoint exerçant une profession d'indépendant» n'étant introduite que dans la première régression).

Dans certains cas de corrélation statistique forte mais non parfaite, on peut songer à introduire un code croisé isolant les cas correspondant à des « incohérences de statut » plutôt que les deux codes en additif. Ainsi supposons que l'on ait introduit deux variables explicatives « être agriculteur », et « être fils d'agriculteur » qui sont statistiquement assez fortement corrélées. On peut souhaiter éliminer les effets nocifs de cette corrélation en créant le code croisé : agriculteur fils d'agriculteur ; agriculteur non fils d'agriculteur, non agriculteur fils d'agriculteur, non agriculteur fils de non agriculteur. Les coefficients seront d'estimation et d'interprétation plus simples. L'exemple n° 2 montre un exemple de sortie SAS, dans le cas d'un modèle mal spécifié, présentant une corrélation mécanique, ainsi que la sortie après correction par dilution de la maladresse de spécification.

The SAS System

The LOGISTIC Procedure

Exemple 2a:
⇒modèle mal spécifié
STR6=DEP75 (Ville de Paris)

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution
Response Levels: 2
Number of Observations: 7332
Link Function: Logit

Response Profile

Ordered Value	IPOLLU	Count
1	1	1161
2	2	6171

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	6408.974	6288.742	.
SC	6415.874	6440.542	.
-2 LOG L	6406.974	6244.742	162.232 with 21 DF (p=0.0001)
Score	.	.	155.685 with 21 DF (p=0.0001)

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

DEP75 = 1 * STR6

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.3164	0.1453	254.0946	0.0001	.	0.099
DIP2	1	0.0911	0.1070	0.7247	0.3946	0.019617	1.095
DIP3	1	0.0256	0.1095	0.0546	0.8153	0.005497	1.026
DIP4	1	-0.1527	0.1315	1.3484	0.2456	-0.026494	0.858
DIP5	1	-0.0294	0.1065	0.0763	0.7824	-0.007406	0.971
STR2	1	0.5403	0.1163	21.5955	0.0001	0.111049	1.716
STR3	1	0.5960	0.1216	24.0269	0.0001	0.113534	1.815
STR4	1	1.0012	0.1012	97.8644	0.0001	0.250680	2.721
STR5	1	0.8663	0.1281	45.7125	0.0001	0.145174	2.378
STR6	1	1.2180	0.1657	54.0260	0.0001	0.132185	3.380
STL1	1	0.1500	0.0899	2.7866	0.0951	0.040872	1.162
STL3	1	0.1646	0.1022	2.5962	0.1071	0.041759	1.179
AGE1	1	-0.0751	0.1261	0.3551	0.5512	-0.013669	0.928
AGE2	1	0.0213	0.1073	0.0394	0.8426	0.004657	1.022
AGE3	1	0.0167	0.1016	0.0271	0.8693	0.003762	1.017
AGE5	1	-0.1480	0.1120	1.7451	0.1865	-0.028719	0.862
AGE6	1	-0.3306	0.1392	5.6431	0.0175	-0.052484	0.719
REV1	1	-0.1624	0.1383	1.3785	0.2404	-0.024617	0.850
REV2	1	-0.1341	0.1106	1.4710	0.2252	-0.027112	0.874
REV3	1	-0.1518	0.1197	1.6095	0.2046	-0.026899	0.859
REV4	1	-0.0620	0.1048	0.3499	0.5542	-0.012481	0.940
REV6	1	-0.0528	0.0948	0.3098	0.5778	-0.012423	0.949
DEP75	0	0	.	.	.	.	.

The SAS System

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 60.9%	Somers' D = 0.231
Discordant = 37.8%	Gamma = 0.234
Tied = 1.2%	Tau-a = 0.062
(7164531 pairs)	c = 0.615

The SAS System

The LOGISTIC Procedure

Exemple 2b:
Correction par dilution
STR5 regroupe STR5 et STR6

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution
Response Levels: 2
Number of Observations: 7332
Link Function: Logit

Response Profile

Ordered Value	IPOLLU	Count
1	1	1161
2	2	6171

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	6408.974	6288.742	.
SC	6415.874	6440.542	.
-2 LOG L	6406.974	6244.742	162.232 with 21 DF (p=0.0001)
Score	.	.	155.685 with 21 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.3164	0.1453	254.0946	0.0001	.	0.099
DIP2	1	0.0911	0.1070	0.7247	0.3946	0.019617	1.095
DIP3	1	0.0256	0.1095	0.0546	0.8153	0.005497	1.026
DIP4	1	-0.1527	0.1315	1.3484	0.2456	-0.026494	0.858
DIP5	1	-0.0294	0.1065	0.0763	0.7824	-0.007406	0.971
STR2	1	0.5403	0.1163	21.5955	0.0001	0.111049	1.716
STR3	1	0.5960	0.1216	24.0269	0.0001	0.113534	1.815
STR4	1	1.0012	0.1012	97.8644	0.0001	0.250680	2.721
STR5	1	0.8663	0.1281	45.7125	0.0001	0.167385	2.378
STL1	1	0.1500	0.0899	2.7866	0.0951	0.040872	1.162
STL3	1	0.1646	0.1022	2.5962	0.1071	0.041759	1.179
AGE1	1	-0.0751	0.1261	0.3551	0.5512	-0.013669	0.928
AGE2	1	0.0213	0.1073	0.0394	0.8426	0.004657	1.022
AGE3	1	0.0167	0.1016	0.0271	0.8693	0.003762	1.017
AGE5	1	-0.1480	0.1120	1.7451	0.1865	-0.028719	0.862
AGE6	1	-0.3306	0.1392	5.6431	0.0175	-0.052484	0.719
REV1	1	-0.1624	0.1383	1.3785	0.2404	-0.024617	0.850
REV2	1	-0.1341	0.1106	1.4710	0.2252	-0.027112	0.874
REV3	1	-0.1518	0.1197	1.6095	0.2046	-0.026899	0.859
REV4	1	-0.0620	0.1048	0.3499	0.5542	-0.012481	0.940
REV6	1	-0.0528	0.0948	0.3098	0.5778	-0.012423	0.949
DEP75	1	0.3517	0.1649	4.5488	0.0329	0.038167	1.421

Association of Predicted Probabilities and Observed Responses

Concordant = 60.9%	Somers' D = 0.231
Discordant = 37.8%	Gamma = 0.234
Tied = 1.2%	Tau-a = 0.062
(7164531 pairs)	c = 0.615

• Les défauts d'additivité :

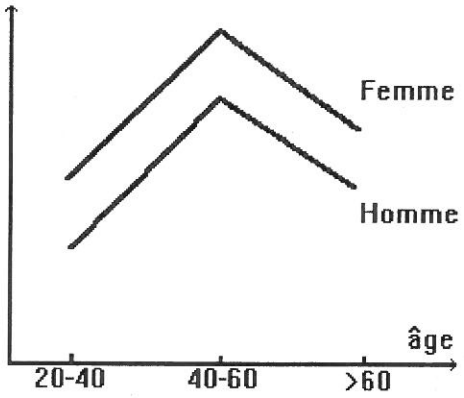
Le modèle latent spécifié sous la forme $Z = \sum_j X_j \beta_j + \varepsilon$ est un modèle additif : l'unicité du coefficient β_j signifie que l'on suppose que la variable X_j agit de la même façon quelles que soient les configurations prises par les variables $X_k, k \neq j$.

Exemple : on étudie le fait de faire de la couture.

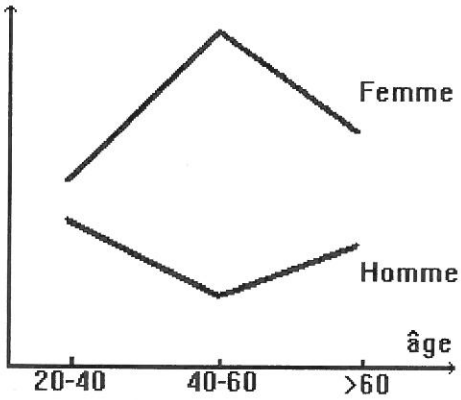
Si l'on écrit le modèle sous la forme:

$$Y = a + bI_{femme} + c_1 I_{âge<40} + c_3 I_{âge>60} + \varepsilon$$

cela revient à supposer que l'on a des effets ayant la forme suivante :



Or on sait que les hommes ne font jamais de couture, quel que soit leur âge. On aurait donc dû chercher à mesurer des profils ayant l'allure suivante :



Ceci ne peut se faire qu'en introduisant de nouvelles variables et en postulant un modèle avec interaction entre l'âge et le sexe soit un modèle de la forme :

$$Y = a + c_1 I_{<40} I_H + c_3 I_{>60} I_H + d_1 I_{<40} I_F + d_2 I_{40-60} I_F + d_3 I_{>60} I_F + \varepsilon$$

où la produit des variables indicatrices correspond à la conjonction d'une tranche d'âge et d'un genre donné. Le modèle comporte alors 6 coefficients à estimer au lieu de 4.

La tentation est grande de rejeter a priori tout modèle spécifié additivement sous le prétexte que la réalité, complexe, ne saurait être approchée d'une façon aussi schématique. Le danger est alors de multiplier les effets d'interaction. Même avec des effectifs importants, on arrive vite à faire croître de façon rédhibitoire

le nombre des coefficients à estimer. Il semble plus adapté de se contenter d'une approximation additive, sauf dans le cas où il est manifeste que le phénomène est différent d'une strate à l'autre (comme dans l'exemple de la couture). Le remède, dans ce cas, sera plus fréquemment de faire deux régressions indépendantes, une sur chacune des strates, plutôt que d'introduire un effet d'interaction limité à un endroit bien particulier du modèle : on étudiera alors de façon indépendante la couture chez les hommes et la couture chez les femmes, en permettant à tous les coefficients de toutes les variables d'être différents dans les 2 régressions.

• Le choix de la situation de référence : les mêmes conseils peuvent être donnés que dans le cas à une seule dimension. On ajoutera seulement que le souci « esthétique » conduit ici à privilégier un choix cohérent des caractéristiques de la situation de référence, afin de reconstituer les diverses facettes du portrait robot d'un parfois hypothétique français « moyen ».

d. les variables omises

Il arrive (et c'est d'ailleurs sans doute le cas pour chaque étude) que les données dont on dispose ne permettent pas d'introduire toutes les dimensions théoriquement pertinentes, même sous forme d'indicateurs indirects. On est alors dans le cas d'un modèle mal spécifié, avec variables omises. On sait que ceci, par le jeu des corrélations entre la variable omise et les variables présentes, est susceptible de biaiser l'estimation des coefficients relatifs à ces variables présentes.

Les considérations précédentes sur les corrélations entre variables présentes peuvent être étendues, avec le même constat rassurant : l'expérience montre que ce n'est que lorsque la variable omise est très corrélée avec une variable présente, ce qui se repère aisément lors de la phase préalable de réflexion sur la spécification du modèle, que l'on court un risque de mauvaise estimation. Dans la plupart des cas, l'estimation des coefficients des variables présentes est suffisamment robuste pour que ces problèmes de variables omises ne fassent pas courir le danger d'un commentaire erroné.

e. Pondérer ou ne pas pondérer : that's the question !

Quelle que soit l'enquête étudiée, se pose le problème de la pondération, au minimum correction de structure, de faible ampleur, destinée à corriger l'échantillon des non réponses différentielles et extrapolation à la population globale, au maximum correction en structure de très forte ampleur, destinée à corriger des non réponses et de taux de sondage initiaux très fortement inégaux, et extrapolation.

Les avis divergent sur l'opportunité de pondérer. Traditionnellement les économistes recommandent de ne pas pondérer car « on étudie des comportements » et le fait que M. Dupont ait un poids égal à 2 ne lui accorde pas d'importance plus grande qu'à M. Durand dont le poids vaut 1.

Cette recommandation est justifiée par le fait qu'en règle générale il n'y aurait pas de relation entre la façon dont les poids sont déterminés et le phénomène analysé, ce qui est loin d'être le cas lorsqu'on surpondère les riches dans une enquête sur le patrimoine ou les grosses firmes dans une enquête salariale.

Cependant, la plupart des discours (parfois enflammés) se réfèrent plus fréquemment à des propos de circonstance ou à des positions de principe qu'à une véritable réflexion méthodologique. Avant de poursuivre, deux remarques permettent d'y voir plus clair :

↳ Le modèle LOGIT (mais pas le PROBIT) possède la propriété que les estimateurs des paramètres de pente (c'est-à-dire, des paramètres relatifs aux variables explicatives) sont invariants à une surreprésentation fondée sur la variable expliquée. Seule la constante du modèle est affectée par la surreprésentation.

C'est typiquement le cas des études médicales : lorsque Y représente le fait d'être atteint d'une maladie, les hôpitaux peuvent sélectionner indépendamment dans deux populations : une population atteinte (Y=1) et une population non atteinte (Y=0). Si la maladie est rare, cela correspond à un fort suréchantillonnage des sujets malades.

Plus précisément, on a le théorème suivant :

Théorème :

Soit D l'événement {un membre de la population est malade}.

Soit $\pi(x) = P(D/x)$ la probabilité d'être malade conditionnellement à x dans la population.

On suppose que $\pi(x)$ suit un modèle logistique dans la population, i.e. $\pi(x) = \dfrac{1}{1+\exp(-\alpha - \beta x)}$

Soit S l'événement {l'individu est échantillonné}

Soit $\rho_0 = P(S/D)$ et $\rho_1 = P(S/D^c)$

Alors $P(D/S,x)$ suit également une loi logistique, avec le même paramètre d'effet β mais avec une ordonnée à l'origine $\alpha^* = \alpha + \log(\dfrac{\rho_0}{\rho_1})$

En effet, par la formule de Bayes $\mu(x) = P(D/S,x) = \dfrac{P(S/D,x)P(D/x)}{P(S/x)}$

Mais $P(S/D,x) = \rho_0, P(D/x) = \pi(x),$

$P(S/x) = P(S/D,x)P(D/x) + P(S/D^c,x)P(D^c/x) = \rho_0\pi(x) + \rho_1(1-\pi(x))$

D'où $\mu(x) = \dfrac{\rho_0\pi(x)}{\rho_0\pi(x) + \rho_1(1-\pi(x))} = \dfrac{\rho_0 \exp(\alpha+\beta x)}{\rho_0 \exp(\alpha+\beta x) + \rho_1} = \dfrac{\frac{\rho_0}{\rho_1}\exp(\alpha+\beta x)}{\frac{\rho_0}{\rho_1}\exp(\alpha+\beta x) + 1}$

Ce qui achève la démonstration.

Autrement dit, dans ce cas précis, il n'est pas nécessaire d'avoir un échantillon bien pondéré pour estimer correctement les paramètres relatifs aux variables explicatives..

Finalement, si le modèle logistique est vrai dans la population, peu importe que l'on surreprésente ou sous-représente dans le tirage certaines modalités de la variable expliquée.

↳ Surreprésentation de certaines modalités de la variable explicative

Dans ce cas, les estimateurs du maximum de vraisemblance pondéré et non pondéré sont différents, car la vraisemblance dépend des effectifs des cases, mais on peut vérifier que dans le cas où on n'a **qu'une seule** dimension explicative discrète, les estimateurs du MV sont les mêmes que l'on pondère ou pas. Il faut alors voir l'estimateur pondéré et l'estimateur non pondéré comme deux versions, qui diffèrent en échantillon fini, mais représentent asymptotiquement les mêmes coefficients des variables explicatives.

Quelques conseils :

Ces deux remarques rendent les considérations sur la pondération ou non des modèles assez relatives. Cependant, les cas concrets peuvent ne pas coïncider exactement avec la situation précédente. En particulier, la sélection des échantillons se fait rarement, voire jamais, dans les enquêtes courantes de l'INSEE, sur la variable expliquée uniquement. Les procédures de redressement de la non-réponse conduisent toujours, en pratique, à des poids qui dépendent de nombreuses variables explicatives ; ces variables n'ont aucune raison d'être toutes présentes dans les modèles que l'on cherche à estimer. Dans un autre ordre d'idée, les praticiens qui utilisent les résultats du modèle pour estimer des probabilités d'émergence du phénomène étudié au sein des diverses strates de population (cf. infra) souhaitent en général retomber au plus près sur les probabilités effectivement constatées dans les strates. Pour cela, ils préfèrent pondérer.

Lorsque les poids sont peu dispersés, les résultats des régressions pondérées et non pondérées au niveau des coefficients sont peu différents et le choix est donc à nouveau de peu d'importance.

Il n'en va pas de même lorsque les poids sont très dispersés même conditionnellement aux variables explicatives (poids allant par exemple de 1 à 40, de 1 à 100 .... Pondérer conduit à la limite à n'estimer le modèle que sur les observations dotées d'un poids élevé, les autres ne jouant plus aucun rôle. Si les comportements des unités à poids élevé sont différents des unités à poids faible, les résultats varient beaucoup entre la version pondérée et la version non pondérée (coefficient changeant de signe par exemple). Une solution peut être de créer des sous-populations au sein desquelles les poids sont peu dispersés et de faire des régressions indépendantes sur chacune des sous-populations, régressions élémentaires pour lesquelles on se trouve ramené au cas précédent, où le choix de pondérer on non a peu d'importance.

Un autre type de difficulté, très pratique cette fois, surgit quand on pondère, au niveau des écarts-types des coefficients et donc des tests de significativité. Il suffit pour s'en convaincre de réaliser deux estimations de la même équation, la première sans pondérer, la seconde en utilisant une pondération uniforme égale à 10 000. les écarts-types sont divisés par 100, alors même que la précision de l'estimation n'a en rien évolué entre les deux variantes. Du point de vue informatique, toutse passe, lorsqu'on pondère, comme si on créait un fichier fictif où chaque individu se verrait doté d'un nombre de jumeaux égal à son poids. Le nuage des points fictifs à étudier est donc constitué par un très grand nombre de points agrégés en petits tas. On conçoit bien que si cette situation était effectivement observée dans la réalité, cela se traduirait par une estimation très précise, d'autant plus précise que le nombre de sosies est grand. On conçoit tout aussi bien que lorsque l'existence de ces classes est le résultat artificiel d'une opération de pondération, le gain de précision n'est qu'un mirage, contre lequel il faut se prémunir.

Pour cela, il importe de toujours utiliser des **pondérations normalisées** de moyenne 1, en divisant la variable de pondération par sa moyenne calculée sur l'ensemble du fichier. Les tests obtenus sont alors utilisables, du moins en première approximation.

A l'exemple n° 3, le lecteur trouvera l'exemple tiré de l'exploitation de la première vague du panel européen des ménages. La régression non pondérée est celle de l'exemple 1, elle n'est donc pas reprise. La régression 3a utilise la pondération brute présente dans le fichier (redressement en structure et extrapolation), la régression 3b utilise la pondération normalisée ramenée à une variable de moyenne égale à 1.

The SAS System

The LOGISTIC Procedure

Exemple 3a
Pondération non normalisée

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution
Response Levels: 2
Number of Observations: 7332
Weight Variable: PONDM
Sum of Weights: 22804144.184
Link Function: Logit

Response Profile

Ordered Value	IPOLLU	Count	Total Weight
1	1	1161	3688498
2	2	6171	19115646

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	20184188	19833424	.
SC	20184195	19833569	.
-2 LOG L	20184186	19833382	350804.56 with 20 DF (p=0.0001)
Score	.	.	347947.94 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.0546	0.00249	683170.282	0.0001	.	0.128
DIP2	1	0.0212	0.00189	125.5123	0.0001	0.255771	1.021
DIP3	1	-0.0498	0.00197	639.6019	0.0001	-0.587310	0.951
DIP4	1	-0.1903	0.00235	6562.2578	0.0001	-1.831404	0.827
DIP5	1	0.0429	0.00187	527.3661	0.0001	0.606142	1.044
STR2	1	0.1778	0.00194	8382.6693	0.0001	1.985481	1.195
STR3	1	0.2416	0.00199	14698.7915	0.0001	2.561458	1.273
STR4	1	0.6354	0.00156	166229.685	0.0001	8.846158	1.888
STR5	1	0.4672	0.00200	54362.3762	0.0001	4.638090	1.596
STL1	1	0.3107	0.00158	38645.5117	0.0001	4.734686	1.364
STL3	1	0.2002	0.00184	11859.9869	0.0001	2.838159	1.222
AGE1	1	0.00142	0.00223	0.4061	0.5240	0.014368	1.001
AGE2	1	0.0340	0.00192	312.1219	0.0001	0.412655	1.035
AGE3	1	0.0366	0.00183	398.2221	0.0001	0.454191	1.037
AGE5	1	-0.1231	0.00201	3734.4483	0.0001	-1.321614	0.884
AGE6	1	-0.3482	0.00230	22950.1225	0.0001	-3.395872	0.706
REV1	1	-0.2858	0.00254	12704.8548	0.0001	-2.356063	0.751
REV2	1	-0.1726	0.00198	7618.2953	0.0001	-1.918919	0.841
REV3	1	-0.1546	0.00214	5222.3580	0.0001	-1.503295	0.857
REV4	1	-0.0584	0.00186	985.0882	0.0001	-0.652953	0.943
REV6	1	0.0231	0.00165	197.2463	0.0001	0.309412	1.023

Association of Predicted Probabilities and Observed Responses

Concordant = 59.0%	Somers' D = 0.193
Discordant = 39.7%	Gamma = 0.196
Tied = 1.3%	Tau-a = 0.051
(7164531 pairs)	c = 0.596

The SAS System

The LOGISTIC Procedure

Exemple 3b
Pondération normalisée

Data Set: WORK.CODIF
Response Variable: IPOLLU Pollution
Response Levels: 2
Number of Observations: 7332
Weight Variable: PONDMC
Sum of Weights: 7356.1755433
Link Function: Logit

Response Profile

Ordered Value	IPOLLU	Count	Total Weight
1	1	1161	1189.8382
2	2	6171	6166.3374

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	6513.028	6439.865	.
SC	6519.928	6584.765	.
-2 LOG L	6511.028	6397.865	113.163 with 20 DF (p=0.0001)
Score	.	.	112.241 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.0546	0.1384	220.3775	0.0001	.	0.128
DIP2	1	0.0212	0.1052	0.0405	0.8405	0.004594	1.021
DIP3	1	-0.0498	0.1097	0.2063	0.6497	-0.010548	0.951
DIP4	1	-0.1903	0.1308	2.1169	0.1457	-0.032893	0.827
DIP5	1	0.0429	0.1040	0.1701	0.6800	0.010887	1.044
STR2	1	0.1778	0.1081	2.7041	0.1001	0.035660	1.195
STR3	1	0.2416	0.1109	4.7415	0.0294	0.046005	1.273
STR4	1	0.6354	0.0868	53.6225	0.0001	0.158882	1.888
STR5	1	0.4672	0.1116	17.5363	0.0001	0.083303	1.596
STL1	1	0.3107	0.0880	12.4663	0.0004	0.085037	1.364
STL3	1	0.2002	0.1024	3.8258	0.0505	0.050975	1.222
AGE1	1	0.00142	0.1243	0.0001	0.9909	0.000258	1.001
AGE2	1	0.0340	0.1070	0.1007	0.7510	0.007411	1.035
AGE3	1	0.0366	0.1020	0.1285	0.7200	0.008158	1.037
AGE5	1	-0.1231	0.1122	1.2047	0.2724	-0.023737	0.884
AGE6	1	-0.3482	0.1280	7.4033	0.0065	-0.060992	0.706
REV1	1	-0.2858	0.1412	4.0983	0.0429	-0.042316	0.751
REV2	1	-0.1726	0.1101	2.4575	0.1170	-0.034465	0.841
REV3	1	-0.1546	0.1191	1.6846	0.1943	-0.027000	0.857
REV4	1	-0.0584	0.1036	0.3178	0.5730	-0.011727	0.943
REV6	1	0.0231	0.0917	0.0636	0.8009	0.005557	1.023

Association of Predicted Probabilities and Observed Responses

Concordant = 59.0%	Somers' D = 0.193
Discordant = 39.7%	Gamma = 0.196
Tied = 1.3%	Tau-a = 0.051
(7164531 pairs)	c = 0.596

f. L'endogénéité

Les problèmes liés à l'endogénéité ou la simultanéité se produisent, comme dans le cas des modèles à variables quantitatives, dès lors que l'on cherche à estimer séparément l'une des équations d'un modèle à équations simultanées se présentant sous forme structurelle. Dans ce cas, la variable expliquée dépend des variables explicatives exogènes X_1 mais aussi de la variable Y_2 suspectée d'endogénéité:

$$Y_1^* = Y_2\delta_1 + X_1\beta_1 + u_1$$

La résolution du système d'équations simultanées permet de mettre en évidence la forme réduite du système bivarié. Supposons que celle-ci s'écrive :

$$Y_1^* = X_1\pi_1 + v_1 \qquad (1)$$
$$Y_2 = X_2\pi_2 + v_2 \qquad (2)$$

avec (v_1, v_2) suivant une loi normale bidimensionnelle.

On suppose ici que la variable Y_2 est quantitative. Moyennant une redéfinition des paramètres, on peut monter que conditionnellement à Y_2, Y_1^* peut s'écrire:

$$Y_1^* = Y_2\delta + X_1\beta + v_2\alpha + u_1^*$$

où u_1^* suit une loi normale conditionnellement à v_2.

Une approche en termes de maximisation de la vraisemblance conditionnelle permet de tester l'endogénéité. Elle consiste à estimer l'équation (2) par les moindres carrés ordinaires. Les variables explicatives peuvent ou non comporter les variables de l'équation (1) mais il doit y avoir au moins une variable instrumentale supplémentaire. On dispose d'un estimateur $\hat{v}_2$ qu'il suffit d'introduire dans l'équation précédente pour obtenir des valeurs non biaisées de β et δ. Si le coefficient α n'est pas significatif, alors on rejette l'hypothèse d'endogénéité. Notons que cette procédure ne s'applique que dans le cas d'un modèle PROBIT, puisque les résidus doivent être normaux.

2. La lecture des résultats

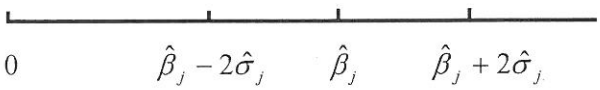
Ici encore les façons de faire divergent d'un statisticien à l'autre, en particulier sur la place à accorder à la notion de significativité globale d'une dimension explicative, et sur la façon de classer par ordre d'importance les diverses dimensions introduites et donc de répondre à une question du type : si on devait ne garder que deux ou trois dimensions explicatives, lesquelles garderait-on ?

Il nous semble que la façon de procéder la plus simple consiste à accorder le maximum d'importance à la significativité des coefficients, pour chacune des variables prise individuellement.

a. significativité des coefficients

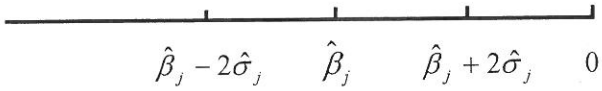
Le test le plus aisé à utiliser est la statistique de Student (ou son carré, la statistique de Wald $\hat{\beta}_j^2 / \hat{V}\hat{\beta}_j = \hat{\beta}_j^2 / \hat{\sigma}_j^2$) (cf partie « test »)

• Si $\hat{\beta}_j / \hat{\sigma}_j > 2$, le coefficient β_j est significativement positif au seuil de 5 % : on a en effet la configuration suivante:



le vrai β_j ayant 95 % de chances d'être entre $\hat{\beta}_j - 2\hat{\sigma}_j$ et $\hat{\beta}_j + 2\hat{\sigma}_j$, intervalle entièrement situé dans l'ensemble des nombres réels positifs.

• Si $\hat{\beta}_j / \hat{\sigma}_j < -2$, le coefficient β_j est significativement négatif : β_j a 95 % de chances d'être entre $\hat{\beta}_j - 2\hat{\sigma}_j$ et $\hat{\beta}_j + 2\hat{\sigma}_j$, intervalle situé dans l'ensemble des nombres réels négatifs puisqu'on a la configuration suivante :



Rq 1 : On peut être moins ambitieux et accepter une notion de significativité à des seuils supérieurs à 5 %. On acceptera alors d'interpréter des coefficients pour lesquels le Student, en valeur absolue est de d'ordre de 1,8, 1,6 voire moins. Ceci sera souvent nécessaire dans les études basées sur des enquêtes à petit échantillon.

Rq 2 : La proc logistic imprime la statistique de Wald, pour la publication il peut être souhaitable de transformer le fichier de sortie pour imprimer la statistique de Student.

b. L'interprétation des coefficients en termes de probabilité

• Cas où toutes les variables explicatives sont des variables 0,1

A partir du taux de pénétration de la pratique dans la situation de référence et des coefficients des diverses variables, on peut donner une interprétation quantitative des résultats obtenus.

Le taux de pénétration de la pratique dans la situation de référence vaut:

$$\Pi_0 = \frac{1}{1+\exp(-\beta_0)} = F(\beta_0)$$

où β_0 est la constante estimée par le modèle.

Pour un individu qui ne dévie de la situation de référence que par la variable X_j le taux de pratique vaut :

$$\Pi = \frac{1}{1+\exp(-\beta_0 - \beta_j)} = F(\beta_0 + \beta_j)$$

Pour un individu qui est affecté de plusieurs déviations $X_j, j \in J$ par rapport à la situation de référence, le taux de pratique vaut:

$$\Pi = \frac{1}{1+\exp(-\beta_0 - \sum_{j\in J} \beta_j)} = F(\beta_0 + \sum_{j\in J} \beta_j)$$

La variation de probabilité entraînée par l'ensemble de déviations est donc:

$$\Delta\Pi = \frac{1}{1+\exp(-\beta_0 - \sum_{j\in J} \beta_j)} - \frac{1}{1+\exp(-\beta_0)}$$

On voit donc qu'il n' a pas additivité des effets de déviations dans l'espace des probabilités, alors qu'il y a additivité dans l'espace des paramètres, et ce à cause de la non linéarité de la fonction:

$$F(x) = \frac{1}{1+\exp(-x)}$$

Sous les hypothèses ici retenues , les déviations par rapport à la situation de référence ne sont pas de petites variations : on saute de tranche d'âge, de tranche de revenu, de CS etc...

On ne peut donc utiliser de formules approchées.

• Cas d'une variable explicative continue.

On peut alors envisager le cas d'une **petite** déviation par rapport à la situation initiale.

Supposons ainsi que l'on s'écarte du revenu initial R_0 de dR. La variation de probabilité s'écrit au premier ordre :

$$d\pi = \beta\pi_0(1-\pi_0)dR$$

où β est le coefficient de la variable revenu et π_0 la probabilité d'observer la pratique dans la situation initiale.

Rq 1 : Lorsqu'il y a suffisamment d'effectif dans la situation de référence, on peut calculer le vrai taux de pratique observé au lieu de l'estimer comme on vient de le faire, et utiliser cette valeur dans les formules précédentes. Certains y voient quelques avantages, en particulier celui d'éviter une discordance entre le taux estimé et le vrai taux, qu'un utilisateur des données peut être tenté de calculer, pour cette situation de référence, à partir de tabulations classiques. On peut cependant s'interroger sur l'aspect « bricolé » d'une telle pratique.

Rq 2 : Lorsqu'on compare des régressions ayant les mêmes variables explicatives mais portant sur deux (ou plusieurs) pratiques différentes (diverses pratiques de loisir, divers biens durables etc...) on peut être tenté de comparer les résultats entre eux.

Une comparaison directe serait trompeuse si on la conduisait à partir des variations de probabilité : une même variation de probabilité n'a pas le même sens selon qu'elle s'applique à une pratique très, très peu répandue ou adoptée par environ la moitié de la population. Le recours à $\Delta\Pi/\Pi$ aurait des inconvénients symétriques. Au total, les deux méthodes conduiraient à des résultats radicalement différents. (voir exemple n°4)

La loi logistique fournit une normalisation qui permet de comparer directement les coefficients de chaque variable d'une régression à l'autre, même si les taux de pratique dans la situation de référence sont très différents. Des coefficients estimés à la même valeur par le modèle correspondent à une intensité équivalente de la disparité étudiée pour les deux pratiques. En termes de probabilité, cette même disparité se traduit par un écart faible lorsque la pratique est répandue dans la moitié de la population, et par des écarts plus forts pour des pratiques peu ou très répandues (il y a symétrie aux deux extrémités de la population).

De fait, la différence logistique est proche d'une échelle multiplicative sur Π si Π est petit, d'une échelle additive si Π est proche de 0,5, et d'une échelle multiplicative sur $1-\Pi$ si Π est voisin de 1. Ceci renvoie à l'intuition correspondant à l'observation des phénomènes de diffusion. Le démarrage et la saturation sont difficiles à obtenir, alors que la diffusion dans les zones médianes est plus aisée.

Deux exemples tirés du panel européen des ménages permettent d'illustrer cette remarque. Etre privé de lave-vaisselle pour des raisons financières (et non par goût) est assez peu fréquent (10 % des ménages sont concernés), alors qu'être privé de résidence secondaire pour les mêmes raisons est plus répandu (près de 40 %

des ménages). Estimons le même modèle logit pour ces deux pratiques. Si on considère la dimension explicative « âge de la personne de référence », l'effet le plus fort est lié à la variable AG6 (avoir plus de 75 ans) ; les ménages les plus âgés se déclarent les moins privés dans les deux cas. Les valeurs des coefficients indiquent que l'effet est particulièrement marqué dans le cas du lave-vaisselle (- 1,15 au lieu de - 0,89 dans le cas de la résidence secondaire).

Si on calcule les différences de probabilité, entre une situation qui diffère de la référence uniquement sous l'aspect « âge » (AG6 au lieu de AG4), et la situation de référence, on obtient :

- pour le lave-vaisselle $\Delta\Pi$ = - 6,66 points (on passe de 10,09 % à 3,43 %)

- pour la résidence secondaire $\Delta\Pi$ = - 18,97 points (on passe de 41,77 % à 22,80 %)

Si on se fiait à ces valeurs, on conclurait à la plus forte importance de l'effet de l'âge dans le cas de la résidence secondaire. Si on se référait enfin à $\Delta\Pi/\Pi_0$, on aurait :

- pour le lave-vaisselle $\Delta\Pi/\Pi_0$ = -0,66 %

- pour la résidence secondaire $\Delta\Pi/\Pi_0$ = - 0,45 %

et donc un constat en sens inverse.

Rq 3 : Juger de la force d'un effet doit également se faire à partir du coefficient, et non pas des différences de probabilité.

On est alors tributaire du choix de la situation de référence ; il suffit de choisir soigneusement une situation de référence pour laquelle le taux de pratique est très faible pour obtenir des $\Delta\Pi/\Pi$ impressionnants et mettre en pleine lumière (artificielle) un effet en réalité assez secondaire.

Rq 4 : Quand on combine des déviations de la situation de référence, rien n'empêche mathématiquement de calculer des probabilités correspondant à des situations impossibles ou loufoques. Le problème se pose donc de savoir si on a le droit de « sortir » de l'échantillon et de reconstituer des probabilités fictives pour des solutions non représentées dans l'échantillon. Ici encore, on se contentera d'une réponse pragmatique : un modèle n'est jamais « presse-bouton », sa mise en oeuvre nécessite toujours un minimum (maximum ?) de bon sens. Ce qui empêchera le statisticien d'utiliser ces combinaisons « monstrueuses ».

The SAS System

The LOGISTIC Procedure *Exemple 4a: Lave-vaisselle*

Data Set: WORK.CODIF
Response Variable: IVAIS Lave vaiss.
Response Levels: 2
Number of Observations: 7344
Link Function: Logit

Response Profile

Ordered Value	IVAIS	Count
1	1	758
2	2	6586

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	4879.693	4459.646	.
SC	4886.595	4604.580	.
-2 LOG L	4877.693	4417.646	460.047 with 20 DF (p=0.0001)
Score	.	.	453.731 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.1873	0.1632	179.5674	0.0001	.	0.112
DIP2	1	-0.3103	0.1254	6.1250	0.0133	-0.066863	0.733
DIP3	1	-0.2842	0.1207	5.5416	0.0186	-0.061031	0.753
DIP4	1	-0.3428	0.1497	5.2472	0.0220	-0.059533	0.710
DIP5	1	-0.3463	0.1261	7.5391	0.0060	-0.087156	0.707
STR2	1	-0.2201	0.1281	2.9543	0.0857	-0.045277	0.802
STR3	1	-0.0174	0.1289	0.0181	0.8929	-0.003306	0.983
STR4	1	0.0823	0.1066	0.5965	0.4399	0.020605	1.086
STR5	1	0.0461	0.1556	0.0879	0.7668	0.007735	1.047
STL1	1	0.3258	0.1069	9.2938	0.0023	0.088767	1.385
STL3	1	-0.3572	0.1385	6.6560	0.0099	-0.090618	0.700
AGE1	1	-0.0386	0.1434	0.0723	0.7881	-0.007020	0.962
AGE2	1	-0.0469	0.1265	0.1375	0.7107	-0.010248	0.954
AGE3	1	-0.0564	0.1237	0.2080	0.6484	-0.012684	0.945
AGE5	1	-0.6107	0.1584	14.8596	0.0001	-0.118568	0.543
AGE6	1	-1.1483	0.2039	31.7114	0.0001	-0.182299	0.317
REV1	1	1.1821	0.1408	70.4817	0.0001	0.179467	3.261
REV2	1	1.0477	0.1245	70.7670	0.0001	0.211768	2.851
REV3	1	0.4907	0.1439	11.6339	0.0006	0.086865	1.633
REV4	1	0.3322	0.1372	5.8647	0.0154	0.066850	1.394
REV6	1	-0.8350	0.1697	24.1961	0.0001	-0.196638	0.434

Association of Predicted Probabilities and Observed Responses

Concordant = 72.3%	Somers' D = 0.455
Discordant = 26.9%	Gamma = 0.458
Tied = 0.8%	Tau-a = 0.084
(4992188 pairs)	c = 0.727

The SAS System

The LOGISTIC Procedure

Exemple 4b:
Résidence secondaire

Data Set: WORK.CODIF
Response Variable: IRS Resid. second.
Response Levels: 2
Number of Observations: 7344
Link Function: Logit

Response Profile

Ordered Value	IRS	Count
1	1	2837
2	2	4507

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	9799.851	9194.478	.
SC	9806.753	9339.412	.
-2 LOG L	9797.851	9152.478	645.373 with 20 DF (p=0.0001)
Score	.	.	606.987 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-0.3319	0.1025	10.4947	0.0012	.	0.718
DIP2	1	-0.1503	0.0825	3.3223	0.0683	-0.032380	0.860
DIP3	1	-0.0765	0.0820	0.8691	0.3512	-0.016427	0.926
DIP4	1	-0.1926	0.0971	3.9350	0.0473	-0.033440	0.825
DIP5	1	-0.3444	0.0822	17.5720	0.0001	-0.086687	0.709
STR2	1	0.1357	0.0776	3.0582	0.0803	0.027904	1.145
STR3	1	0.0673	0.0830	0.6568	0.4177	0.012804	1.070
STR4	1	0.4591	0.0674	46.3801	0.0001	0.114908	1.583
STR5	1	0.3204	0.0931	11.8548	0.0006	0.053721	1.378
STL1	1	0.0237	0.0658	0.1298	0.7186	0.006457	1.024
STL3	1	-0.4751	0.0784	36.7676	0.0001	-0.120513	0.622
AGE1	1	0.3744	0.0967	14.9933	0.0001	0.068174	1.454
AGE2	1	0.5066	0.0820	38.2125	0.0001	0.110667	1.660
AGE3	1	0.3193	0.0778	16.8368	0.0001	0.071817	1.376
AGE5	1	-0.4338	0.0918	22.3493	0.0001	-0.084217	0.648
AGE6	1	-0.8875	0.1172	57.3647	0.0001	-0.140897	0.412
REV1	1	-0.1978	0.1034	3.6578	0.0558	-0.030026	0.821
REV2	1	0.00376	0.0826	0.0021	0.9637	0.000760	1.004
REV3	1	-0.0984	0.0895	1.2104	0.2713	-0.017427	0.906
REV4	1	0.0325	0.0801	0.1647	0.6849	0.006540	1.033
REV6	1	-0.5514	0.0761	52.5060	0.0001	-0.129864	0.576

Association of Predicted Probabilities and Observed Responses

Concordant = 66.5%	Somers' D = 0.335
Discordant = 33.0%	Gamma = 0.337
Tied = 0.5%	Tau-a = 0.159
(12786359 pairs)	c = 0.667

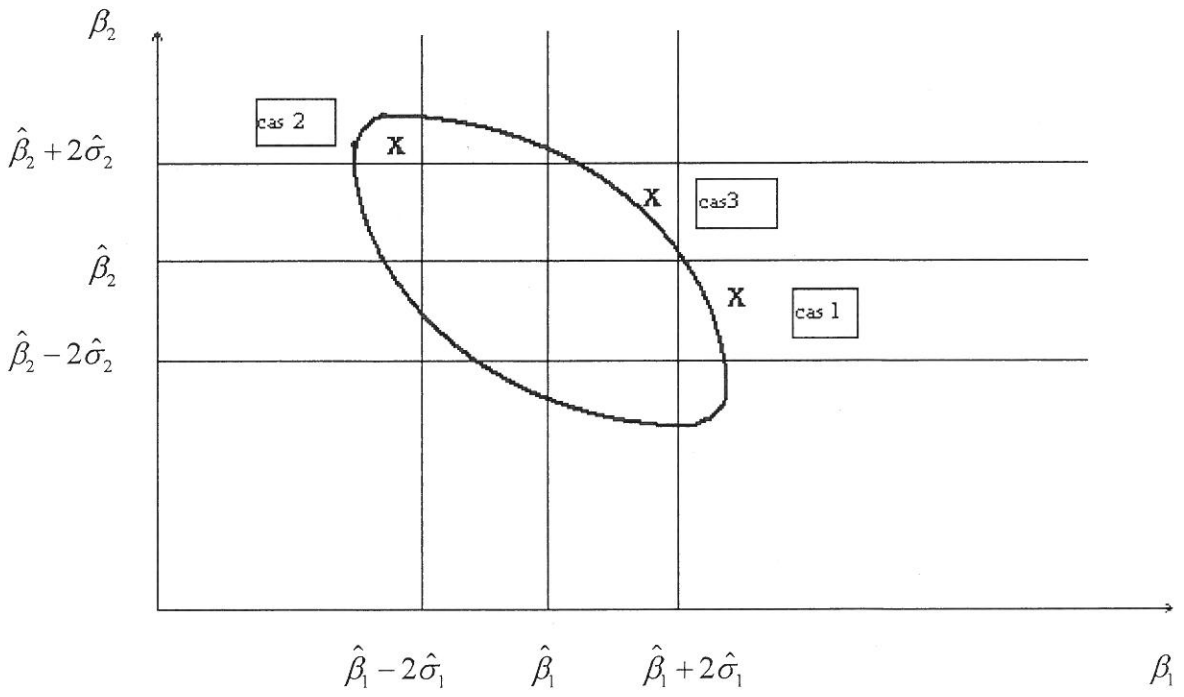
c. significativité globale d'une dimension explicative.

Différents moyens se présentent pour juger de la significativité globale d'une dimension explicative, sans qu'aucun n'apparaisse supérieur à l'autre:

• Une première méthode fondée sur le pouvoir discriminant de la dimension dans la population considérée consiste à réaliser un test sur le caractère explicatif de toutes les variables introduites pour représenter cette dimension. On teste ainsi l'égalité à zéro de tous les coefficients introduits. Le test du rapport de vraisemblance se conduit aisément : on estime dans un premier temps le modèle avec les variables représentant la dimension et dans un deuxième temps le modèle sans ces variables. L'opposé du double des écarts des logarithmes des vraisemblances suit alors un χ^2 à autant de degrés de liberté que l'on a introduit de variables pour représenter la dimension. Plus la statistique obtenue s'éloigne du seuil de significativité, plus on considère que la dimension apporte de l'information. Un moyen plus aisé en termes de programmation consiste à utiliser l'instruction `Test` de la `Proc Logistic`, afin de tester la nullité de tous les coefficients correspondant aux variables représentant la dimension. La procédure fournit alors un test de Wald, asymptotiquement équivalent au test du rapport de vraisemblance. Notons qu'à l'instar de la `Proc GLM`, la `Proc Probit` fournit directement la statistique de Wald correspondant à la dimension dès lors que l'instruction `Class` est utilisée.

• Une seconde méthode consiste à examiner les différents coefficients eux-mêmes et en évaluer la significativité indépendamment de celle des autres coefficients de la même dimension.

Aucune solution n'est en soi préférable à l'autre, et les deux techniques apparaissent comme complémentaires lors de l'élaboration de l'interprétation. Le test de Wald peut refuser la non significativité de la dimension alors même qu'aucun coefficient pris individuellement n'est significatif. De même, la significativité de la dimension peut être rejetée alors qu'un ou plusieurs coefficients sont significatifs pris individuellement. Ce dernier cas est beaucoup plus rare. Par exemple, dans le cas de deux coefficients, on peut montrer que le rectangle correspondant au produit des intervalles de confiance au niveau α pour les deux paramètres est une région de confiance à (1-2α) % pour le paramètre bidimensionnel. Cela signifie que si un des deux paramètres est significatif à 5 %, il y a moins de 10 % de chances que le paramètre bidimensionnel ne soit pas significatif.



La figure montre les intervalles de confiance associés aux deux paramètres ainsi que l'ellipsoïde de confiance (pour le même niveau de confiance) associé au paramètre bidimensionnel. Les croix représentent diverses positions alternatives du point (0,0). On a illustré trois cas :
Cas 1 : les deux paramètres sont significatifs, la dimension aussi.
Cas 2 : les paramètres sont significatifs, la dimension ne l'est pas..
Cas 3 : les deux paramètres ne sont pas significatifs, mais la dimension globale l'est.

Dans le commentaire du rôle joué par la dimension, il convient donc d'utiliser les deux approches, notamment lorsque les coefficients sont peu significatifs. Les tests de coefficients isolés peuvent parfois apparaitre « pessimistes » ! Ainsi, dans le cas où la dimension explicative est continue ordonnée et que l'on obtient des coefficients qui dessinent un profil régulier conforme à ce que la théorie permet de prévoir, on peut prendre argument d'une significativité de la dimension au sens du test de Wald et risquer de commenter le signe d'un coefficient alors même que le Student correspondant est faible (de l'ordre de 1 par exemple).

d. peut-on classer les diverses dimensions explicatives par ordre d'importance (« puissance explicative ») ?

Plutôt que de juger de la seule significativité d'une dimension explicative, on peut souhaiter aller plus loin et établir un ordre entre les dimensions, des plus explicatives aux moins informatives. Les problèmes rencontrés sont de même nature que précédemment, les différences de classement entre les méthodes provenant du traitement différent des strates peu nombreuses mais très atypiques. A nouveau, le classement est affaire de point de vue :

• On peut envisager de comparer les variations de pouvoir explicatif global du modèle entraînées par l'adjonction de chaque dimension explicative prise une à une (on introduit toutes les variables correspondantes et on calcule le gain de pouvoir explicatif). La variable la plus explicative serait celle qui entraîne la plus forte variation. Cette solution présente l'inconvénient d'un coût informatique élevé. En outre, on peut trouver peu judicieux d'estimer des modèles dont on sait pertinemment qu'ils sont mal spécifiés.

• On peut aussi pour chaque dimension, calculer la valeur absolue de l'écart entre le plus fort coefficient significativement positif (ou 0 s'il n'y en a pas) et le plus fort coefficient significativement négatif (ou 0 s'il n'y en a pas) et classer les diverses dimensions explicatives selon ces valeurs. La rigueur scientifique de cette pratique n'est pas absolue : en particulier si on change la codification de la dimension explicative (tranches plus fines par exemple), on peut modifier le classement. Toutefois, ici encore, si le modèle est convenablement spécifié, cette méthode fournit un moyen fiable de séparer les dimensions de 1ère importance de celles de 2ème niveau. Il ne faut toutefois pas en attendre un véritable classement.

Afin de pallier les insuffisances de ces méthodes, on peut conseiller :

• de faire un grand usage de variantes (tester la robustesse du classement face à des changements dans la façon dont les dimensions explicatives sont représentées). D'une façon générale l'usage de variantes est un bon moyen de se rendre compte du degré auquel on peut ajouter foi aux résultats du modèle : on peut recourir à des variantes sur la définition de l'échantillon, sur le degré auquel on introduit de l'interactivité, sur la codification etc...

• de comparer avec d'autres méthodes statistiques. Comme toute méthode d'analyse statistique, l'analyse Logit repose sur des hypothèses plus ou moins implicites (en particulier celle déjà discutée concernant l'additivité). D'autres méthodes comme l'analyse de correspondances ou la segmentation reposent sur d'autres combinaisons d'hypothèses. Si ces méthodes confirment les résultats du Logit (ou du moins ne les infirment pas), on sera tout à fait à l'aise pour commenter les résultats, sans état d'âme particulier. C'est ainsi que les expériences antérieures ont montré que le classement des dimensions explicatives effectué comme on vient de le décrire, recoupait presque systématiquement les conclusions que fournissent des analyses de segmentation réalisées à partir des mêmes variables.

e. les coefficients égaux à $\pm\infty$

Ceci se produira chaque fois que pour une strate définie par une variable personne ou tout le monde est concerné par la pratique étudiée. La probabilité Π_j à estimer vaut alors 0 ou 1. Vu la forme de la fonction exponentielle,

$$\frac{1}{1+\exp(-\beta_0-\beta_j)}$$

cette probabilité ne peut valoir 0 que si $\beta_j = -\infty$ (alors $\exp(-\beta_0-\beta_j) = +\infty$).

De même, elle ne peut valoir 1 que si $\beta_j = +\infty$. Ceci risque de se produire dès que l'on a une strate d'effectif faible et que l'on étudie une pratique très peu ou très répandue. Les algorithmes utilisés dans la procédure d'estimation ont des conditions d'arrêt qui empêchent de partir à l'infini.

La proc logistic, face à une telle occurrence se comporte de la façon suivante :

• 1er cas : c'est le cas le plus fréquent ; le listing ne comporte que la page « simple statistics for Explanatory variables » et la mention « Convergence was not attained in 25 iterations » (voir exemple n° 2 pour un fac similé de la sortie d'un message de ce type, obtenu dans le cadre de l'analyse des problèmes de colinéarité). Le nombre d'itérations prévu par défaut (ie 25) ne suffit pas. Il faut relancer en modifiant ce nombre d'itérations en rajoutant dans la carte spécifiant le modèle, l'instruction `Maxiter = n` (n > 25) indiquant le nombre maximum d'itérations à réaliser. Afin de réduire la fréquence des cas où il faut relancer, on peut conseiller de procéder à l'estimation, dès le début, en introduisant l'instruction `Maxiter = 50` : si le modèle converge rapidement cela n'a aucune influence, dans le cas contraire on évitera la plupart du temps d'avoir à relancer, et on se retrouvera dès le premier passage dans le 2ème cas.

• 2ème cas : celui-ci ne se produit plus que très rarement à partir de la version 6.08 de SAS ; le modèle a convergé, mais la page de résultats présente des coefficients estimés très forts. On est alors dans le cas d'un $\beta_j = \pm\infty$. Il conviendra donc, dans la publication, de remplacer le coefficient apparaissant dans le listing par $+\infty$ ou $-\infty$ selon le cas. Les autres coefficients sont estimés correctement : tout se passe comme si l'on travaillait sur le sous-fichier obtenu en supprimant les observations correspondant à la strate pour laquelle le coefficient est infini ; ce n'est que pour le calcul des statistiques de test global que des différences apparaissent.

La situation se présente de façon analogue, mais un peu plus complexe lorsque c'est dans une des strates composant la situation de référence que la pratique étudiée est absente (ou au contraire omniprésente).

Deux possibilités sont utilisables pour détecter ce type de situation. La première consiste à réaliser une analyse des fréquences respectives des deux modalités de la variable expliquée pour l'ensemble des variables explicatives. Une fréquence nulle ou égale à 100 % pour l'une des variables entraîne une absence de convergence. De fait, une telle analyse devrait être réalisée systématiquement avant de mettre en oeuvre une `Proc Logistic`, afin de mieux comprendre les résultats. Un autre moyen consiste à utiliser l'option `Itprint` de l'instruction `Model`. La valeur de tous les coefficients sera imprimée à chaque itération. Il est alors aisé de détecter les coefficients qui tendent vers l'infini.

Dans le cas de coefficients infinis, deux solutions se présentent:
- soit exclure la sous-population concernée. On travaille sur un sous-échantillon.
- soit regrouper cette sous-population avec une strate voisine, de sorte que la fréquence de la pratique cesse d'être nulle ou égale à 100 %. On conserve alors l'échantillon complet, mais l'analyse perd en finesse.

f. derniers problèmes

Une fois tous les problèmes précédents résolus, une fois obtenu un modèle convergent, sans aucune anomalie visible, il se peut qu'un dernier piège guette le statisticien imprudent. L'estimation peut être en fait fragile, car ne reposant que sur des effectifs très réduits. Un signe révélateur peut éveiller les soupçons : c'est l'existence parmi les écarts-types estimés de valeurs très fortes (de l'ordre de 100 par exemple, alors que les autres sont de l'ordre de 1 ou inférieurs). C'est par exemple le cas lorsque la population de référence est trop peu nombreuse dans l'une de ses dimensions. On peut facilement repérer ce type de situation en croisant préalablement la variable expliquée avec les variables explicatives.

La prudence conseille alors de s'assurer de la robustesse des résultats, en modifiant, sur le point douteux, la spécification du modèle. Une fois encore on ne saurait trop recommander l'usage systématique de variantes (de spécification, d'échantillon, de population de référence ...).

3 - La publication des résultats

Un bref survol des publications récentes comportant des modèles logit suffit à prouver qu'il n'y a pas encore de standards de présentation. De l'information minimale, ne faisant apparaître que les signes des coefficients significativement différents de 0 avec éventuellement indication des coefficients les plus « marqués » à la présentation complète de tous les coefficients avec les écarts-types correspondant, qui seule peut donner satisfaction aux économètres, la gamme des solutions choisies est assez étendue (voir exemples 7).

Les responsables de publications « grand public » répugnent souvent à noyer le lecteur sous un torrent de chiffres, surtout s'il s'agit de coefficients dont le caractère abstrait ne peut être levé sans un important effort de la part du lecteur pour assimiler la théorie économétrique. Ils conseilleront alors, soit de publier uniquement un tableau de + et -, soit de publier les probabilités ou les différences de probabilité (en absolu ou relatif) entraînées par les déviations de la situation de référence.

La première solution a l'inconvénient de réduire très fortement la quantité d'information transmise, puisqu'on perd toute indication de l'intensité des effets. Les efforts pour tourner cette difficulté ne sont que des palliatifs, car les solutions disponibles ne sont pas irréprochables : on peut choisir de faire apparaître par un graphisme spécial (en gras, avec +++, ...) les effets pour lesquels la statistique de Student est la plus forte (mais ne confond-on pas alors dans un seul chiffre importance de l'effet à proprement parler et précision de son estimation ?) ou ceux pour lesquels les coefficients sont les plus forts (en étant significativement différents de 0 évidemment). Le choix soulève alors le même type de difficultés que celles évoquées pour la détermination de la puissance explicative. De plus, où s'arrêter : doit-on observer une règle de conduite du type « on attire l'attention sur les 3 (ou 4 ou 2) coefficients les plus forts » systématiquement, ou doit-on se laisser guider par les ruptures dans la distribution des coefficients et avoir un nombre variable, déterminé au cas par cas, d'effets mis en évidence ? On voit que la place laissée à la subjectivité de l'auteur est grande (trop ?).

La seconde solution, comme on l'a vu plus haut, risque d'induire le lecteur en erreur en l'incitant à effectuer d'une régression à l'autre des comparaisons illicites. Elles est particulièrement dangereuse lorsque l'on étudie dans le même article plusieurs pratiques très diversement répandues.

Les impératifs techniques, d'autre part, peuvent rendre difficile la publication d'un modèle ayant une centaine de variables explicatives.

Peut-on, ou non, publier un extrait du modèle centré sur un effet particulier ?

On ne peut qu'être réticent à l'idée d'une telle pratique : les résultats sont conditionnels à la spécification du modèle et il importe que le lecteur puisse se faire une idée des qualités et limites de l'ensemble des variables retenues. D'autre part ce n'est que face à l'ensemble des coefficients que le lecteur peut « étalonner » son regard et juger par lui-même s'il s'agit d'un effet de première importance ou non.

Il faut donc publier le modèle intégralement. Quand cela est vraiment impossible, il faut au moins donner la liste des variables introduites et éviter toute publication où n'apparaîtraient que les coefficients relatifs à une seule dimension explicative.

A plusieurs reprises, dans le cours de cette note, on a distingué les dimensions explicatives de nature continues de celles plus « qualitatives ».

Dans le cas des variables continues ordonnées, on ne saurait trop recommander la publication des profils dessinés par les divers coefficients relatifs à cette dimension (voir exemple n° 8).

Exemples n°7

Effets des caractéristiques du ménage sur le statut d'occupation *

	Locataire	Accédant récent	Accédant ancien	Propriétaire
Chef de ménage :				
— de 30 ans ou moins	+	+	—	—
— *de 30 à 45 ans*				
— de 46 à 65 ans	—	—	—	+
— de 66 à 75 ans	—	—	—	+
— de plus de 75 ans	—	—	—	+
Ménage composé :				
— d'un individu			—	
— d'un couple seul	—			
— *d'un couple avec un enfant*				
— d'un couple avec deux enfants			+	—
— d'un couple avec trois enfants ou plus	+	—	+	—
Marié depuis plus de deux ans				
— marié depuis moins de deux ans	+		—	—
Revenu du ménage :				
— inférieur à 35 000 F	+	—	—	—
— de 35 000 à 55 000 F	+	—	—	
— *de 55 000 à 80 000 F*				
— de 80 000 à 110 000 F	—		+	
— supérieur à 110 000 F	—		+	
Chef de ménage :				
— inactif		—	—	+
— agriculteur		—	—	+
— patron				+
— ouvrier	+		+	
— *employé*				
— cadre moyen	—		+	
— cadre supérieur			+	
Lieu de résidence :				
— commune rurale	—	+		+
— *commune urbaine hors agglomération parisienne*				
— agglomération parisienne	—			
— ville de Paris			—	+
Type d'habitat :				
— *habitat collectif*				
— habitat individuel	—	+	+	+

** Ces effets sont étudiés toutes choses égales par ailleurs (annexe p. 30). Pour chaque caractéristique, la situation de référence par rapport à laquelle sont étudiés les effets est indiquée en italique. L'absence de signe indique que l'effet n'est pas statistiquement significatif; le signe renforcé (+ ou —) souligne les effets les plus marqués.*

Source : Economie et Statistique, n° 161, décembre 1983, p.24

Tableau 1 : effets des caractéristiques socio-démographiques sur la propriété du logement principal

Caractéristiques socio-démographiques		Propriété du logement principal
Statut matrimonial x âge de l'homme		
cohabitation juvénile	x moins de 30 ans	—
	30-35 ans	(–)
union libre avant mariage	x moins de 30 ans	—
	30 ans et plus	
mariage	x moins de 30 ans	——
	30-35 ans	*
	35-45 ans	+
	45 ans et plus	+++
cohabitation non juvénile	x moins de 35 ans	–
	35 ans et plus	(–)
union libre après mariage	x moins de 45 ans	–
	45 ans et plus	
Urbanisation		
communes rurales		+++
unités urbaines de moins de 100 000 h		*
unités urbaines de 100 000 h et plus		—
banlieue parisienne		—
Paris		——
Revenu du couple		
moins de 75 000 F		—
de 75 000 F à 100 000 F		*
de 100 000 F à 130 000 F		
de 130 000 F à 200 000 F		+++
200 000 F et plus		+++
Profession de l'homme		
agriculteur		
commerçant artisan		(+)
cadre		
profession intermédiaire		
employé		
ouvrier		*
inactif		
Profession du père de l'homme		
agriculteur		(+)
indépendant		
cadre		
employé		
ouvrier		*
inactif		
Niveau d'éducation de l'homme		
sans diplôme		—
niveau intermédiaire		*
niveau Bac ou études supérieures		

La situation de référence par rapport à laquelle sont étudiés les effets est repérée par un *, l'absence de signe indique que l'effet n'est pas statistiquement significatif, les signes entre parenthèses, simples, doubles ou triples, indiquent les effets, des moins marqués aux plus marqués.

Source : Economie et Prévision, n° 91, 1989, p.111

Les facteurs explicatifs de la production domestique : résultats d'un modèle "toutes choses égales"

Variables introduites dans le modèle	Etre multi-pratiquant	Faire un vêtement en couture	Faire des conserves	Faire des réparations sur des app. mén. ou la voiture	Faire des petits trvx de bricol. ds le log.	Semer, planter des légumes	S'occuper de plantes d'appart.
Age du chef de ménage							
1- moins de 30 ans	0,7	- 0,3	- 0,5	0,7	0,5	- 0,7	- 0,3
2- de 30 à 39 ans		- 0,3	- 0,2	0,3	0,3	- 0,4	- 0,2
3- de 40 à 49 ans*							
4- de 50 à 59 ans	0,3		0,7			0,4	
5- de 60 à 69 ans			0,5	- 0,4	- 0,5	0,6	
6- de 70 à 79 ans	- 1,7	- 0,4	0,3	- 1,2	- 1,3	0,3	- 0,4
7- 80 ans et plus	- 1,8	- 1,3	- 0,5	- 1,5	- 1,7	- 0,5	- 0,7
Taille du ménage							
0- homme seul	- 3,2	- 2,1	- 1,8	- 0,5	- 0,6	- 0,5	- 2,0
1- femme seule	- 3,0		- 0,8	- 2,7	- 1,6	- 1,4	
2- 2 personnes	- 0,4			- 0,2			
3- 3 personnes*							
4- 4 personnes	0,3	0,3	0,2	0,2	0,2		0,5
5- 5 personnes	0,6	0,4	0,4				
6- 6 personnes et plus	0,9	0,7	0,5				
Catégorie socioprofessionnelle du chef de ménage							
1- agriculteur		- 0,4	0,6	- 0,7	- 0,7	0,9	
2- artisan, com., chef d'entr.	- 0,5			- 0,5	- 0,4	- 0,5	
3- cadre supérieur		0,3	- 0,3	- 0,3		- 0,8	
4- profession intermédiaire			- 0,3		0,3	- 0,5	0,4
5- employé			- 0,2	- 0,3	- 0,3	- 0,4	0,4
6- ouvrier*							
Diplôme du chef de ménage							
0- aucun diplôme	- 0,6	- 0,4	- 0,4	- 0,4	- 0,4		
1- CEP ou assimilable	- 0,7	- 0,2	- 0,2	- 0,3	- 0,3		
2- CAP *							
3- BEPC						- 0,3	
4- Bac technique ou BP							
5- Bac général					- 0,3		
6- niveau supérieur au bac							
Catégorie de commune							
1- commune rurale			0,4		- 0,3	0,5	- 0,2
2- UU de moins de 20 000 hab*							
3- UU de 20 000 à 100 000 hab			- 0,2			- 0,3	
4- UU de plus de 100 000 hab	- 0,6		- 0,7			- 0,4	
5- "grande couronne"	- 1,5		- 1,5			- 0,7	
6- "petite couronne"			- 1,1		0,2	- 0,5	- 0,4
7- Paris			- 0,8		0,6	- 0,9	- 0,8
Statut d'occupation du logt							
1- propriétaire ou accédant	0,6		0,5	0,2	0,4	0,6	0,4
2- locataire*							
3- logé gratuitement			0,3			0,3	
Type d'immeuble							
1- maison individuelle	2,2	0,2	0,8	0,5	0,6	1,7	0,2
2- immeuble de 2 logements	1,8		0,5		0,5	1,1	
3- immeuble plus de 2 logts*							
Aide ménagère							
0- aucune aide			0,4	0,4	0,7	0,5	
1- employé de maison*							
2- autre type d'aide			- 0,6				
Ressources du ménage							
1- moins de 30 000 F	- 2,0	- 0,7	- 0,6	- 1,0	- 0,6	- 0,5	- 0,8
2- de 30 000 à 49 999 F	- 1,2			- 0,8	- 0,6		- 0,6
3- de 50 000 à 74 999 F	- 0,6			- 0,3	- 0,3		- 0,2
4- de 75 000 à 99 999 F*							
5- de 100 000 à 129 999 F			- 0,2				
6- de 130 000 à 199 999 F					0,3		0,3
7- de 200 000 à 299 999 F			- 0,5			- 0,3	0,5
8- 300 000 F et plus	- 1,2	- 0,5	- 0,6	- 0,4		- 0,4	
9- revenus non déclarés	- 1,0	- 0,6	- 0,7	- 0,5	- 0,4	- 0,6	

*Modalité choisie pour la situation de référence

Lecture : Pour une modalité donnée d'une variable donnée, et pour chacune des activités (y.c. la multi-pratique), le coefficient est d'autant plus élevé que les ménages dans cette situation ont une pratique de cette activité plus fréquente que ceux qui sont dans une situation choisie comme référence. Par exemple, les foyers de 6 personnes et plus, toutes choses égales d'ailleurs, une plus grande chance d'être "multi-pratiquant" (coefficient + 0,9) que ne l'ont les foyers de 3 personnes, pris ici comme situation de référence.

Source : Insee - Enquête "modes de vie" 1988-1989

Source : INSEE-Première, n° 109, octobre 1990

*Analyse des difficultés de recrutement**

	Toutes catégories			Ouvriers non qualifiés			Ouvriers qualifiés			Techniciens et cadres		
	Coeff	Stud	EPro	Coeff	Stud	EPro	Coeff	Stud	EPro	Coeff	Stud	EPro
Constante	- 1,11	6,10	0	- 3,51	9,90	0	- 1,67	8,09	0	- 1,87	8,98	0
Goulots de main-d'œuvre												
Ni goulot, ni gêne, possibilité de produire davantage avec plus de personnel	0,40	2,68	0,08	ns			ns			0,36	2,08	0,05
Ni goulot, ni gêne, pas de possibilité de produire davantage	Réf		0	Réf		0	Réf		0	Réf		0
Gêne (avec ou sans possibilité de produire davantage avec plus de personnel)	3,03	11,24	0,62	1,72	5,32	0,11	2,28	10,57	0,49	1,29	5,93	0,22
Goulot, possibilité de produire davantage avec plus de personnel	2,06	8,47	0,47	1,16	3,45	0,06	1,94	8,76	0,41	0,56	2,43	0,08
Goulot, pas de possibilité de produire davantage	2,36	5,49	0,53	1,94	4,20	0,14	1,62	4,80	0,33	1,13	3,28	0,19
Capacités de production												
Plus que suffisantes	ns			0,66	2,16	0,03	ns			ns		
Normales, pas de goulot d'équipement	Réf		0	Réf		0	Réf		0	Réf		0
Normales, existence de goulots d'équipement	ns			ns			ns			ns		
Insuffisantes, pas de goulot d'équipement	0,72	3,60	0,16	1,04	4,21	0,05	0,39	2,04	0,06	0,49	2,59	0,07
Insuffisantes, existence de goulots d'équipement	0,53	2,50	0,11	ns			0,56	2,58	0,09	0,42	1,88	0,06
Evolutions des effectifs												
Augmentation passée, augmentation prévue	1,01	4,02	0,23	1,19	4,14	0,06	0,42	1,82	0,06	0,85	3,99	0,13
Augmentation passée, stabilité prévue	0,43	2,33	0,09	ns			0,32	1,73	0,05	0,33	1,78	0,04
Augmentation passée, diminution prévue	1,00	1,90	0,22	ns			1,16	2,29	0,22	ns		
Stabilité passée, augmentation prévue	0,49	1,75	0,10	ns			ns			ns		
Stabilité passée, stabilité prévue	Réf		0	Réf		0	Réf		0	Réf		0
Stabilité passée, diminution prévue	ns			ns			ns			ns		
Diminution passée (augmentation, stabilité ou diminution prévue)	ns			ns			ns			ns		
Evolutions de la demande												
Augmentation passée, augmentation prévue	0,52	2,04	0,11	0,78	2,32	0,03	ns			ns		
Augmentation passée, stabilité prévue	0,51	2,70	0,11	0,60	2,13	0,02	ns			ns		
Augmentation passée, diminution prévue	ns			ns			ns			ns		
Stabilité passée, (augmentation, stabilité ou diminution prévue)	Réf		0	Réf		0	Réf		0	Réf		0
Diminution passée (augmentation, stabilité ou diminution prévue)	ns			ns			ns			ns		
Evolutions de la production												
Augmentation passée, augmentation prévue	ns			ns			ns			ns		
Augmentation passée, stabilité prévue	ns			ns			ns			ns		
Augmentation passée, diminution prévue	ns			ns			ns			ns		
Stabilité passée (augmentation, stabilité ou diminution prévue)	Réf		0	Réf		0	Réf		0	Réf		0
Diminution passée (augmentation, stabilité ou diminution prévue)	ns			- 1,17	2,30	- 0,02	ns			ns		
Opinion sur les stocks												
Supérieurs à la normale	0,50	2,80	0,11	ns			ns			0,57	3,05	0,08
Normaux	Réf		0	Réf		0	Réf		0	Réf		0
Inférieurs à la normale	ns			ns			ns			ns		
Jamais de stocks	0,39	2,52	0,08	ns			0,31	1,90	0,05	ns		
Taille												
De 10 à 100 salariés	- 0,29	2,19	- 0,05	0,80	4,00	0,03	- 0,32	2,30	- 0,04	- 0,88	5,89	- 0,07
De 100 à 500 salariés	Réf		0	Réf		0	Réf		0	Réf		0
Plus de 500 salariés	ns			- 1,67	3,43	- 0,02	- 0,66	3,55	- 0,07	ns		
Secteur												
Biens intermédiaires	Réf		0	Réf		0	Réf		0	Réf		0
Biens d'équipement professionnel	ns			- 0,50	2,11	- 0,01	ns			0,35	2,29	0,05
Automobile, matériel de transport terrestre	ns			ns			ns			ns		
Biens de consommation et d'équipement ménager	- 0,68	5,04	- 0,10	ns			- 0,49	3,26	- 0,06	- 0,45	2,96	- 0,04

Champ : entreprises de l'industrie manufacturière.

Source : enquête trimestrielle de conjoncture sur la situation et les perspectives dans l'industrie d'octobre 1989.

Source : Economie et Statistique, n° 234, juillet-août 1990, p.9

Modèle PROBIT de la détention d'actifs

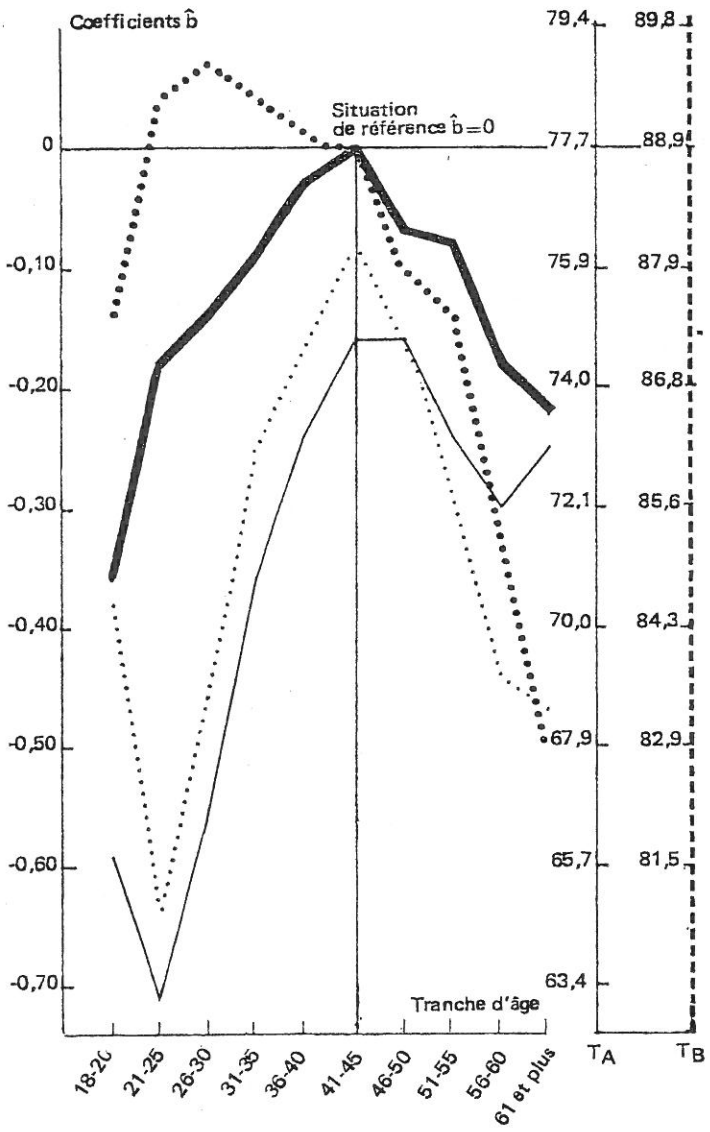
Variables explicatives	1	2	3	4	5	6	7
Constante	0,045	0,435	−1,706	−1,870	−2,801	−2,201	−2,851
	(0,348)	(0,168)	(0,215)	(0,118)	(0,249)	(0,106)	(0,111)
Patrimoine (10E-7)	8,079	0,151	1,053	0,270	1,032	0,372	2,019
	1,990)	(0,264)	(0,225)	(0,120)	(0,248)	(0,121)	(0,068)
(Patrimoine)^2 (10E-14)	−0,768	0,004	−0,342	−0,027	−0,253	−0,036	−0,213
	(12,372)	(0,143)	(0,076)	(0,017)	(0,086)	(0,017)	(0,011)
Revenu (10E-6)	6,179	0,468	0,633	1,301	0,477	0,960	0,871
	(0,955)	(0,183)	(0,209)	(0,183)	(0,199)	(0,104)	(0,177)
Age (10E-1)	0,336	−0,062	0,042	0,445	0,274	0,608	0,680
	(0,136)	(0,068)	(0,087)	(0,037)	(0,093)	(0,030)	(0,029)
(Age)^2 (10E-2)	−0,040	0,009	0,005	−0,054	−0,017	−0,073	−0,049
	(0,012)	(0,006)	(0,008)	(0,004)	(0,009)	(0,003)	(0,002 6)
Héritage (héritier = 1)	0,121	0,111	0,184	0,155	0,161	0,071	0,371
	(0,094)	(0,045)	(0,047)	(0,041)	(0,055)	(0,042)	(0,040)
Donation versée	0,102	0,107	−0,175	−0,044	0,124	0,111	−0,052
	(0,147)	(0,088)	(0,092)	(0,090)	(0,103)	(0,094)	(0,076)
Sit. Matr. (Marié = 1)	0,048	0,116	−0,157	0,171	0,021	0,289	0,579
	(0,102)	(0,053)	(0,060)	(0,051)	(0,073)	(0,050)	(0,052)
Nombre d'enfants	0,001	−0,017	−0,025	−0,047	−0,065	0,008	−0,011
	(0,019)	(0,012)	(0,014)	(0,013)	(0,018)	(0,013)	(0,012)
Femme active	0,230	0,157	0,086	0,106	0,198	0,268	0,125
	(0,099)	(0,051)	(0,057)	(0,049)	(0,068)	(0,047)	(0,050)
Entrepreneur Individuel	0,023	−0,177	0,045	0,201	−0,076	0,203	0,056
	(0,122)	(0,059)	(0,064)	(0,054)	(0,079)	(0,056)	(0,054)
Exploitant agricole	0,308	−0,217	0,499	0,341	0,270	−0,077	0,312
	(0,135)	(0,070)	(0,072)	(0,071)	(0,089)	(0,075)	(0,073)
Niveau d'éducation 1	0,295	0,315	0,077	0,235	0,245	0,145	0,239
	(0,082)	(0,052)	(0,061)	(0,055)	(0,080)	(0,054)	(0,052)
Niveau d'éducation 2	0,872	0,352	0,168	0,341	0,356	0,242	0,288
	(0,172)	(0,069)	(0,078)	(0,067)	(0,098)	(0,065)	(0,067)
Niveau d'éducation 3	1,046	0,320	0,824	0,421	0,302	0,126	0,119
	(0,246)	(0,082)	(0,095)	(0,079)	(0,115)	(0,077)	(0,081)
Niveau d'éducation 4	1,318	0,335	0,237	0,526	0,111	0,072	0,045
	(0,333)	(0,076)	(0,085)	(0,073)	(0,120)	(0,074)	(0,074)
Niveau d'éducation 5	1,320	0,343	0,217	0,708	0,353	0,198	0,104
	(0,418)	(0,082)	(0,085)	(0,073)	(0,102)	(0,070)	(0,075)
Nombre de détenteurs	5 373	4 455	791	1 896	442	1 881	3 083
Khi2 (17)	495,50	148,82	212,33	655,99	175,96	721,70	1 384,42

Source : Annales d'Economie et de Statistique, n° 17, janvier/mars 1990, p. 26
(ce tableau n'est qu'un extrait d'un tableau plus grand)

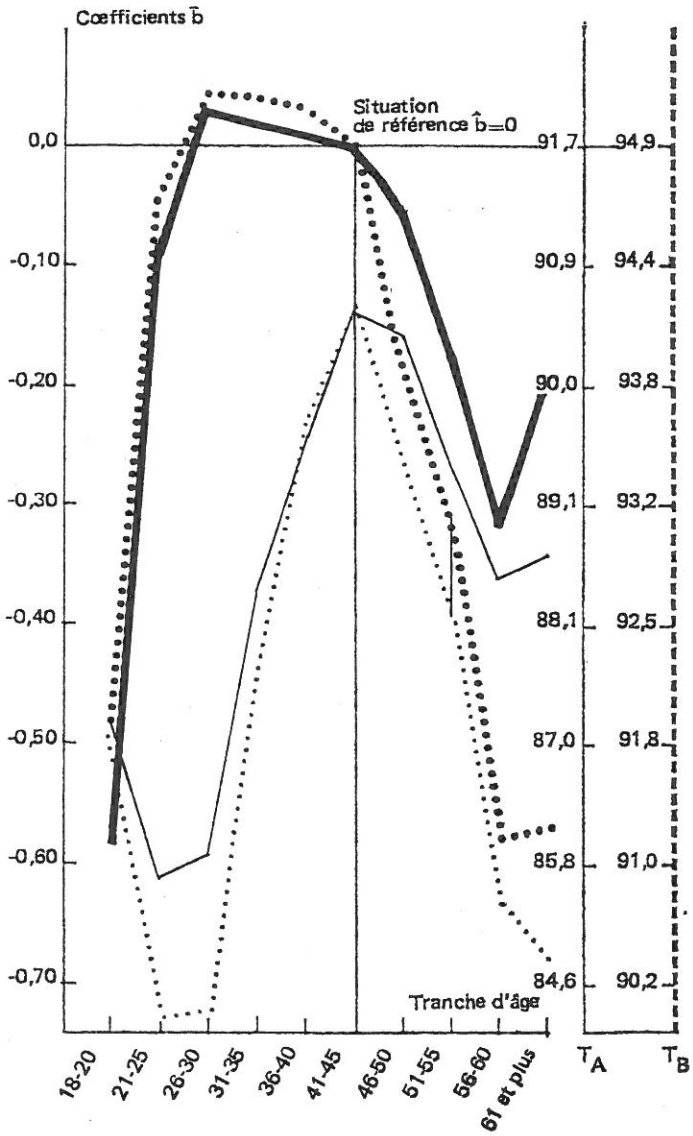
Exemple n°8

Disparités de présence par sexe et âge *

I.A. Ouvriers



I.B. Salariés non ouvriers



Taux de présence estimés «toutes choses égales par ailleurs»
T_A : % de présents tout le mois
T_B : % de présents ou d'absents de moins de cinq jours dans le mois

Variable expliquée
A. Présence tout le mois
Hommes ▬▬
Femmes ——
B. Absence de moins de cinq jours
Hommes • • • • •
Femmes · · · · ·

* *L'interprétation simultanée des coefficients* $\hat{b}$ *et des taux de présence estimés est explicitée dans l'encadré p. 20.*

Source : Economie et de Statistique, n° 176, avril 1985, p. 17

X Quelques problèmes économétriques souvent ignorés

1. L'hétéroscédasticité

Dans le cas du modèle logit, on a vu qu'on pouvait supposer l'existence d'une variable latente Y_n^* de la forme $Y_n^* = X_n\beta + u_n$ où $Eu_n = 0$ et $Vu_n = \sigma^2$ (indépendante de n).

Il y a hétéroscédasticité si la variance du résidu u_n n'est plus indépendante de n, c'est-à-dire si l'on a :

$$Y_n^* = X_n\beta + u_n \text{ avec } Eu_n = 0 \text{ et } Vu_n = \sigma_n^2$$

Dans ce cas la méthode de l'estimation par le maximum de vraisemblance n'est pas convergente.

Davidson et Mac Kinnon (1984) ont proposé un test dans le cas particulier où σ_n^2 serait de la forme :

$$\sigma_n^2 = \frac{\pi^2}{3}\exp(2Z_n\gamma)$$

où Z_n est un vecteur de k_1 variables supposées avoir de l'influence sur σ_n^2 et γ un vecteur de paramètres.

On teste alors l'hypothèse nulle d'homoscédasticité, H_0 : $\gamma = 0$ qui correspond à $\sigma^2 = \pi^2/3$ pour tout n (cas du logit) contre l'hypothèse alternative de ce cas particulier d'hétéroscédasticité H_a : $\gamma \neq 0$.

On voit, en divisant $Y_n^* = X_n\beta + u_n$ par $\exp(Z_n\gamma)$, qu'on se ramène dans le cas de l'hypothèse alternative à une sorte de modèle logit plus général, non linéaire par rapport aux paramètres β et γ, et tel que :

$$p_n = \frac{1}{1+\exp(-X_n\beta / \exp(Z_n\gamma))}$$

Davidson et Mac Kinnon utilisent le test du score. La statistique du score suit asymptotiquement dans l'hypothèse nulle la loi d'un χ^2 à k_1 degrés de libertés (où k_1 est le nombre de variables de Z) et vaut :

$$\text{Score} = S_1' I^{11} S_1$$
$$\text{où } S_1 = \sum_n (Y_n - \hat{p}_n)(-X_n\hat{\beta})Z_n'$$
$$\text{et } I^{11} = \sum_n \left[(Y_n - \hat{p}_n)(X_n\hat{\beta})\right]^2 Z_n' Z_n$$

$\hat{p}_n$ et $\hat{\beta}$ sont les estimateurs de p_n et β dans le cas de l'hypothèse nulle.

On accepte l'hypothèse nulle (cas du logit homoscédastique habituel) si la statistique du score est inférieure à un certain seuil.

2. L'asymétrie de la distribution des perturbations

Dans le cas du modèle LOGIT, les perturbations ont pour distribution la loi logistique.

La loi logistique est un cas particulier de la loi de Burr pour laquelle :

$$p_n = P[Y_n = 1] = \frac{1}{[1 + \exp(-X_n\beta)]^{\alpha}}$$

Dans le cas de $\alpha = 1$, on retrouve le modèle logit. Pour $\alpha < 1$ la distribution est plus « épaisse » sur la gauche, pour $\alpha > 1$ elle est plus « épaisse » sur la droite.

On effectue le test du score, pour tester l'hypothèse nulle H_0 : $\alpha = 1$ contre l'hypothèse alternative : $\alpha \neq 1$.

La statistique du score suit asymptotiquement dans l'hypothèse nulle la loi du χ^2 à un degré de liberté.

Elle est de la forme :

$$\text{Score} = S_1' I^{11} S_1$$
$$\text{où} \quad S_1 = \sum_n (\log \hat{p}_n)(Y_n - \hat{p}_n)/(1 - \hat{p}_n)$$
$$\text{et} \quad I^{11} = \sum_n [(\log \hat{p}_n)(Y_n - \hat{p}_n)/(1 - \hat{p}_n)]^2$$

où $\hat{p}_n$ est l'estimateur de p_n dans l'hypothèse nulle (logit).

On conclura à la validité de l'hypothèse nulle lorsque la statistique du score sera inférieure à 4.

3 Test de mauvaise spécification

Le test dit « de la matrice d'information » peut être utile à la fois pour tester les cas de mauvaise spécification du modèle (hétéroscédasticité, erreur sur la distribution etc...) et le cas de « variation aléatoire des coefficients » encore appelé de manière générale hétérogénéité. Ce dernier terme signifie que certains coefficients β_j ne sont pas constants pour toutes les observations mais qu'ils varient d'une observation à l'autre.

Le principe de ce test est le suivant :

Si L_n est la log-vraisemblance du modèle, alors :

$$V\frac{\partial L_n}{\partial \beta} = -E\frac{\partial^2 L_n}{\partial \beta \partial \beta'}$$

Si cette égalité n'est pas vérifiée, alors le modèle n'est pas bien spécifié ; la distribution à la base du calcul de L_n n'est pas celle ayant engendré les données.

Michael Lechner (1989) donne quatre statistiques asymptotiquement équivalentes permettant d'effectuer le test dans le cas du logit. La première est due à White (1982), la deuxième à Chesher et Lancaster (1984), les deux dernières à Orme (1988). Ces statistiques sont d'un calcul un peu complexe. (voir tableau joint).

Les statistiques de ORME, plus simples, semblent préférables.

Le test de la matrice d'information

Ce tableau présente quatre statistiques permettant d'effectuer le test de la matrice d'information.

On notera :

IM (White) la statistique proposée par White (1982)

IM (Ch-La) la statistique proposée par Chesher et Lancaster

IM (Orme 1) la première statistique proposée par Orme (1988)

IM (Orme 2) la deuxième statistique proposée par Orme

IM^{White}	$= D(\hat{\beta})V(\hat{\beta})^{-1}D(\hat{\beta})$		
IM^{Ch-La}	$= i'Y(\hat{\beta})\left[Y(\hat{\beta})'Y(\hat{\beta})\right]^{-1}Y(\hat{\beta})'i$		
IM^{Orme1}	$= a(\hat{\beta})'W(\hat{\beta})\left[W(\hat{\beta})'W(\hat{\beta})\right]^{-1}W(\hat{\beta})'a(\hat{\beta})$		
IM^{Orme2}	$= n\left[a(\hat{\beta})'a(\hat{\beta})\right]^{-1} IM^{Orme1}$		
$d_n(\hat{\beta})$	$= vech\left\{\left\{r_n^2 - (1-\hat{p}_n)\hat{p}_n\right\}X_n'X_n\right\}(\hat{\beta})$		q x 1
$g_n(\hat{\beta})$	$= r_n X_n'$		k x 1
$Y(\hat{\beta})$	$= (d_n', g_n')$	n=1,...,N	N x (q+k)
$D(\hat{\beta})$	$= d_n$		q x 1
$dD(\hat{\beta})$	$= \left\{\left[2Y_n(\hat{p}_n^2 - \hat{p}_n) - 4\hat{p}_n^3 + 5\hat{p}_n^2 - \hat{p}_n\right]vech(X_n'X_n)X_n\right\}$		q x k
$V(\hat{\beta})$	$= \left\{\left[d_n - dD(-I^1)^{-1}g_n\right]\left[d_n - dD(-I^1)^{-1}g_n\right]'\right\}$		q x q
$W(\hat{\beta})$	$= \left[Q(\hat{\beta}), Z(\hat{\beta})\right]$		N x (q+k)
$a(\hat{\beta})$	$= (r_n\left[(1-\hat{p}_n)\hat{p}_n\right]^{-1/2})$	n=1,...,N	N x 1
$Q(\hat{\beta})$	$= \left(\left[(1-\hat{p}_n)\hat{p}_n\right]^{-1/2}X_n\right)$	n=1,...,N	N x k
$Z(\hat{\beta})$	$= (vech\left\{\left[(1-\hat{p}_n)\hat{p}_n\right]^{1/2}(1-2\hat{p}_n)(X_n'X_n)\right\})$	n=1,...,N	N x q

N est le nombre d'observations ; k le nombre de variables explicatives.

i est le vecteur N x 1 composé de 1 ; $r_n = Y_n - \hat{p}_n$

vech désigne l'empilement sur une colonne des éléments distincts d'une matrice symétrique

$\hat{p}_n$ est l'estimateur de p_n par le modèle qu'on cherche à tester

$$I^1 = -E\frac{\partial^2 L_n}{\partial\beta\partial\beta'} = \sum(1-\hat{p}_n)\hat{p}_n X_n'X_n \quad \text{(estimation)}$$

Référence : ce tableau est repris de Michael Lechner (1989)

XI Extension au cas d'une variable dépendante polytomique ordonnée

Exemples :

Faire du sport 1. tous les jours
2. une ou plusieurs fois par semaine
3. plus rarement

Partir en vacances 1. chaque année (y compris plusieurs fois) ou presque
2. plus rarement
3. jamais ou presque jamais

La procédure Logistic permet de traiter ce cas mais sous une hypothèse assez restrictive dite d'« égalité des pentes ».

Supposons que Y, la variable à expliquer, puisse prendre un (petit) nombre de valeurs ordonnées, soient $1,\ldots,i,\ldots,k+1$.

Le modèle ajusté par la procédure Logistic est basé sur les probabilités de distribution cumulées, soit celles de réalisation de $Y \leq i$, plutôt que sur les probabilités de réalisation de $Y = i$. (cf brochure SAS)

La forme du modèle est :

$$P(Y \leq i) = F(\alpha_i + X\gamma) \qquad 1 < i < k$$

où $\alpha_1,\ldots,\alpha_k$ sont les k constantes et γ le vecteur des autres paramètres, ceux correspondant aux variables explicatives, qu'on peut appeler « paramètres de pente » (slope parameters). L'interprétation des constantes soulève les mêmes difficultés que dans le cas dichotomique. Si l'analyse du comportement se réfère à une variable latente, être dans tel ou tel état dépend de la position de cette dernière par rapport à différents seuils. Les contraintes d'identification ne permettent pas de calculer la « constante » de la variable latente **et** les différents seuils. Seul l'écart est identifiable. La conséquence est que les paramètres α_i ont un signe opposé aux seuils, et se présentent donc en sortie dans un ordre inversé.

La constante α_i seule changeant avec i, γ restant le même, il s'agit d'un modèle de régression selon des parallèles, autrement dit avec égalité des pentes quand i varie, ce qui est une hypothèse assez forte.

La syntaxe

La syntaxe de la procédure est la même. Une option précisant l'ordre de tri pour les valeurs prises par la variable dépendante peut être indiquée dans l'instruction `Proc Logistic:`

```
Proc Logistic   Order = DATA
                ou Formatted
                ou Internal
```

`Order = Data` signifie que les valeurs de Y sont triées selon leur ordre d'apparition dans la table SAS en entrée.

`Order = Formatted` si le tri se fait selon la valeur formatée.

`Order = Internal` s'il se fait selon la valeur non formatée

Par défaut, `Order = Formatted` s'il il y a un format précisé par l'utilisateur ; sinon l'option par défaut est `Order = Internal`.

Le test de l'hypothèse d'égalité des pentes

Dans les sorties, ce test s'appelle « Score test for the equal slopes assumption » quand `Link = Normit` (modèle PROBIT) ou `Cloglog` (loi de Gompertz). Quand `Link = Logit`, le test s'appelle « Score test for the proportional odds assumption ». Le mode de calcul de ce test est le suivant :

On fait l'hypothèse que le nombre de valeurs prises par la variable dépendante, k + 1, est plus grand que 2. On suppose qu'il y a s variables explicatives dans le modèle.

Soit le modèle :

$$P(Y \leq i) = F(\alpha_i + X\gamma_i)$$

où $i = 1, \ldots k$, Y est la variable dépendante, $\gamma_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{is})$ le vecteur des s paramètres « de pente », et α_i la constante correspondant à la modalité i.

Dans l'hypothèse d'égalité des pentes, on a :

$\gamma_{1m} = \gamma_{2m} = \ldots = \gamma_{km}$ pour tous m=1, ...,s

Soient $\hat{\alpha}_1, \ldots, \hat{\alpha}_k$ et $\hat{\gamma}_1, \ldots, \hat{\gamma}_s$ les estimateurs du maximum de vraisemblance des constantes et des paramètres de pente dans l'hypothèse de l'égalité des pentes. Alors pour tout i, on a :

$$\hat{\beta}_i = (\hat{\alpha}_i, \hat{\gamma}_1, \ldots, \hat{\gamma}_s)'$$

Si $U(\beta)$ désigne le vecteur des dérivées partielles de la log-vraisemblance par rapport à β, et $I(\beta)$ l'information de Fisher, on estime alors la statistique du score $U'(\beta)I^{-1}(\beta)U(\beta)$ au point $\hat{\beta}_0 = (\hat{\beta}_1, \ldots, \hat{\beta}_k)'$.

Sous l'hypothèse $\beta = \beta_0$, la statistique du score tend asymptotiquement vers une distribution du χ^2 à s (k-1) degrés de liberté. elle permet de tester l'hypothèse de l'égalité des pentes, qui peut être considérée comme vérifiée si la statistique du score ne dépasse pas un seuil α.

Conclusion

Comme dans tous les modes d'emploi, la liste des précautions à prendre est longue, le catalogue des « pannes » ou du moins des incidents qui peuvent survenir impressionnant. Mais comme pour la plupart des appareils, l'usage quotidien est en réalité des plus simples. Les modèles logit sont à la fois simples à mettre en oeuvre et très performants ; à l'expérience nous n'avons jamais rencontré de cas où les traditionnels tableaux croisés démentaient les résultats économétriques ou permettaient d'aller plus loin. Bien au contraire un seul passage d'un modèle logit suffit à économiser des dizaines (voire centaines) de tableaux croisés.

Le travail de rédaction d'un article n'en est pas vraiment simplifié pour autant ; finis les paragraphes pour expliquer longuement qu'un effet peut en cacher un autre, que ce que l'on croit être un effet du revenu est peut-être celui du nombre d'enfants. Dès le début les résultats sont toutes choses égales par ailleurs. L'auteur doit trouver ailleurs matière à copie ..., et il doit être vigilant ; il est facile au détour d'un raisonnement, d'avancer une explication qui pour être séduisante n'en est pas moins hors de propos car elle oublie que l'on est toutes choses égales par ailleurs.

La bibliographie jointe démontre, par son volume, que les modèles logit et assimilés ont ces dernières années su séduire un nombre grandissant de statisticiens ... n'en déplaise à ceux pour qui la démarche même consistant à vouloir séparer des effets est un non sens. Mais, sur ce dernier point, le débat « idéologique » reste ouvert.

Bibliographie

Quelques articles assez anciens mais donnant des éléments d'explication sur la modélisation :

Daniel VERGER; « L'achat d'un logement ne va pas sans achats d'équipements », *Economie et Statistique*, N° 161 décembre 1983.

Alain TROGNON, « Modèle de diffusion d'une innovation : l'exemple de la télévision couleur », *Annales de l'INSEE*, n° 29, janvier 1978.

Daniel DEPARDIEU, Stéfan LOLLIVIER, « Les facteurs de l'absentéisme », *Economie et Statistique*, n° 176, avril 1985.

Stéfan LOLLIVIER, Daniel VERGER, « Les comportements en matière d'épargne et de patrimoine », *Economie et Statistique*, n° 202, septembre 1987. (logit polytomique univarié ordonné).

Quelques articles récents :

Luc Arrondel, « Patrimoine des ménages : toujours le logement, mais aussi les actifs de précaution », *Economie et Statistique*, n° 296-297, 1996-6/7.

Luc Arrondel, André MASSON, « Gestion du risque et comportements patrimoniaux », *Economie et Statistique*, n° 296-297, 1996-6/7.

Didier BALSAN, Saïd HANCHANE, Patrick WERQUIN, « Salaire d'efficience et théorie de la recherche d'emploi : la mobilité de l'emploi vers un autre emploi », *Economie et Statistique*, n° 290, 1995-10.

Alice BARTHEZ, Anne LAFFERRE, « Contrats de mariage et régimes matrimoniaux », *Economie et Statistique*, n° 296-297, 1996-6/7.

Pascal BOUYAUX, « Une difficulté d'interprétation de l'approche LOGIT : l'exemple de l'économie des transports », *Economie et Prévision*, n° 91, 1989.

François CLANCHE, « Le confort des logements dessine aussi l'espace social », *Economie et Statistique*, n° 288-289, 1995-8/9.

Olivier CHOQUET, François HERAN, « Quand les élèves jugent les élèves et les lycées », *Economie et Statistique*, n° 293, 1996-3.

Danielle DELL'ERA, Mireille FLOREMONT, Olivier LEFEVRE, Dominique ROUSSEL, « Le défi de l'emploi à Metz et à Nancy », *Economie et Statistique*, n° 294-295, 1996-4/5.

Olivier GALLAND, « Une entrée de plus en plus tardive dans la vie adulte », *Economie et Statistique*, n° 283-284, 1995-3/4.

Pascal GARRIGUES, « Une France un peu plus sportive qu'il y a vingt ans ... grâce aux femmes », *Economie et Statistique*, n° 224, septembre 1989.

Benédicte GALTIER, « Gérer la main d'oeuvre dans la durée : des pratiques différenciées en renouvellement », *Economie et Statistique*, n° 298, 1996-8.

Louis LEVY-GARBOIS, Claude MONTMARQUETTE, « Une étude économétrique de la demande de théâtre sur données individuelles », *Economie et Prévision*, 121, 1995-5.

Stéfan LOLLIVIER, « Activité et arrêt d'activité féminine », *Economie et Statistique*, n° 212, juillet-août 1988.

Stéfan LOLLIVIER, « Activité des femmes mariées et hétérogénéité : estimation sur données de panel », *Annales d'Economie et de Statistique*, n° 39, juillet/août 1995, 93-106.

Sergio PERELMAN,Pierre PESTIEAU, « Les legs volontaires en France : évaluation et explication », *Economie et Prévision*, 100-101, 1991-4/5.

Laurent TOULEMON, H. LERIDON, « Maitrise de la fécondité et appartenance sociale », *Population*, 1, 1992, 1-46.

Louis André VALLET, « L'assimilation scolaire des enfants issus de l'immigration et son interprétation : un examen sur données françaises », *Revue Française de Pédagogie*, 1996.

Pour des références théoriques complètes :

Christian GOURIEROUX, Econométrie des variables qualitatives, *Economica*, 1989, 2ème édition.

Alain AGRESTI, *Categorical data Analysis,* John Wiley & Sons, 1990.

et un survey un peu ancien :

T. AMEMIYA, « Qualitative response models : a survey », *Journal of economic litterature*, vol. XIX, pp. 1483-1536, décembre 1981.

Sur la partie « Quelques problèmes économétriques souvent ignorés » :

Michael LECHNER, « Testing logit models in practice », *Document de travail*, université d'Heidelberg (se le procurer auprès des auteurs de cette note).

H. WHITE, « Maximum Likelihood Estimation of Misspecified Models », *Econometrica* 50 : 1-25, 1982.

A. CHESHER, « Testing for Neglected Heterogeneity », *Econometrica* 52, 865-872, 1984.

T. LANCASTER, « The Covariance Matrix of the Information Matrix Test », *Econometrica* 52 : 1051-1053, 1984.

C. ORME, « The Calculation of the Information Matrix Test for Binary Data Models », *The Manchestel School* 56, 370-376.

Série des Documents de Travail 'Méthodologie Statistique'

9601 : 'Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population'
G. DECAUDIN, J.-C. LABAT

9602 : 'Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises'
N. CARON, P. RAVALET, O. SAUTORY

9603 : 'La procédure FREQ de SAS® - Tests d'indépendance et mesures d'association dans un tableau de contingence'
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : 'Les principales techniques de correction de la non-réponse et les modèles associés'
N. CARON

9605 : 'L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration'
P. RAVALET

9606 : 'L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)'
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : 'Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes'
N. CARON, D. LE BLANC

9701 : 'Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?'
J.C. DEVILLE

9702 : 'Modèles univariés et modèles de durée sur données individuelles'
S. LOLLIVIER

9703 : 'Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises'
N. CARON, J.C. DEVILLE

9704 : 'La faisabilité d'une enquête auprès des ménages
1. au mois d'août. 2. à un rythme hebdomadaire'
C. LAGARENNE, C. THIESSET

9705 : 'Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine'
P. GIRARD

9801 : 'Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS'
K. ATTAL-TOUBERT, D. LADIRAY

9802 : 'Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation'
J.C. DEVILLE

9803 : 'Pour essayer d'en finir avec l'individu Kish'
J.C. DEVILLE

9804 : 'Une nouvelle (encore une !) méthode de tirage à probabilités inégales'
J.C. DEVILLE

9805 : 'Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish'
J.C. DEVILLE

9806 : 'Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE'
N. CARON, J.C. DEVILLE, O. SAUTORY

9807 : 'Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle'
K. ATTAL-TOUBERT, O. SAUTORY

9808 : 'Matrices de mobilité et calcul de la précision associée'
N. CARON, C. CHAMBAZ

9809 : 'Echantillonnage et stratification : une étude empirique des gains de précision'
J. LE GUENNEC

9810 : 'Le Kish : les problèmes de réalisation du tirage et de son extrapolation'
C. BERTHIER, N. CARON, B. NÉROS

9811 : 'Vocabulaire statistique Français - Chinois - Anglais'
LIU Xiaoyue, CUI Bin

9901 : 'Perte de précision liée au tirage d'un ou plusieurs individus Kish'
N. CARON

9902 : 'Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen'
N. CARON

0001 : 'L'économétrie et l'étude des comportements. Présentation et mise en oeuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT)' (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER