CHAPITRE 4-3
MODÈLES À VARIABLE DÉPENDANTE QUALITATIVE

Plan

CHAPITRE 4-3
MODÈLES À VARIABLE DÉPENDANTE QUALITATIVE

4-3.1 Modèles de choix binaire : logit binomial et probit binomial

4-3.1.1 LE PROBLÈME

Nous abordons ici en quelque sorte l'envers de l'analyse de variance. Dans l'analyse de variance, la variable de réaction est continue, alors que les variables stimuli sont discrètes ; ici, la variable de réaction est discrète, alors que les variables stimuli peuvent être continues.

On peut dériver le modèle logit à partir de l'analyse des tables de contingence. Mais dans l'analyse des tables de contingence, toutes les variables de classification ont des rôles symétriques ; dans le modèle logit au contraire, l'une des variables joue le rôle de variable dépendante, et les probabilités d'appartenance à ses différentes catégories sont conditionnées par les autres variables, qui jouent donc le rôle de variables indépendantes. En outre, la possibilité d'avoir des variables continues parmi les variables indépendantes constitue une généralisation par rapport au logit dérivé de l'analyse des tables de contingence.

Considérons d'abord le cas où la variable dépendante est dichotomique, et non pas polytomique : il n'y a que deux possibilités.

$y_i = 0$ ou 1

Exemple : dans une étude sur la mobilité résidentielle des ménages [1],

- $y_i = 0$ si le ménage *i* ne déménage pas
- $y_i = 1$ si le ménage *i* déménage

Or, dans le modèle linéaire standard

$$y_i = \sum_j x_{ij}\beta_j + u_i$$

la valeur de la variable dépendante y_i ne peut pas être limitée aux valeurs 0 et 1 si les termes aléatoires u_i ont une distribution normale (parce qu'une variable normale est continue et que son domaine de variation s'étend de $-\infty$ à $+\infty$).

---

1 Pour un exemple, voir Mongeau (s.d.).

Le premier pas vers la solution de cette difficulté est de considérer que la véritable variable dépendante n'est pas la variable dichotomique *y*, mais la *probabilité* que $y = 1$:

$$\Pr[y_i = 1] = \sum_j x_{ij}\beta_j + u_i$$

Cela laisse cependant subsister deux difficultés :

1. Bien que la nouvelle variable dépendante, $\Pr[y_i = 1]$, soit continue, son domaine de variation est limité à l'intervalle [0, 1], alors que celui du terme aléatoire, si celui-ci a une distribution normale, s'étend de $-\infty$ à $+\infty$.
2. La nouvelle variable dépendante, $\Pr[y_i = 1]$, n'est pas observée ; ce qu'on observe, ce ne sont que les *réalisations* ($y_i = 1$ *ou* $y_i = 0$) : il faudra recourir à une méthode d'estimation qui s'accommode de cela.

4-3.1.2 MODÈLE DE COMPORTEMENT

L'un des fondements théoriques possibles du modèle logit binomial est un modèle de comportement du type stimulus-réaction (*stimulus-response*), fréquemment utilisé, notamment en biologie : les sujets d'une expérience sont soumis à des conditions (stimuli) qui varient d'un sujet à l'autre ; on observe les réactions et on essaie d'estimer la relation entre stimuli et réaction.

Ce modèle peut se présenter en deux parties. La première partie constitue le modèle de réaction du sujet (modèle de comportement) et la seconde, le modèle du stimulus total qui résulte de la combinaison des conditions auxquelles est soumis un sujet.

Première partie : modèle de réaction

Le stimulus total auquel est soumis le sujet n'est pas directement observé. Supposons néanmoins qu'il y ait une variable non observée (« latente ») *w*, qui mesure ce stimulus, c'est-à-dire l'attrait, ou la « désirabilité » d'un choix (un économiste dirait l'« utilité »). Dans l'exemple de l'étude sur la mobilité résidentielle, le ménage *i* déménage si cette variable latente *w* dépasse une certaine valeur critique, un « seuil de réaction » (*S*) :

$y_i = 0$ si $w_i < S$

$y_i = 1$ si $w_i \geq S$

où w_i est la valeur de la variable latente pour le ménage *i*, et *S* est le seuil critique au-delà duquel le ménage déménage. Il est commode de supposer que la variable latente *w* est définie de telle façon que *S* soit zéro, ce qui donne :

$y_i = 0$ si $w_i < 0$

$y_i = 1$ si $w_i \geq 0$

Cette dernière hypothèse n'impose aucune restriction au modèle, puisque la variable *w* est une variable ordinale, dont le zéro est arbitraire.

Deuxième partie : modèle du stimulus total

La valeur de la variable latente *w* (non observée) est déterminée par un certain nombre de facteurs, mesurés par des variables indépendantes appropriées (les *x*). Dans une étude sur la mobilité résidentielle, par exemple, les variables indépendantes pourraient comprendre les caractéristiques du ménage, les caractéristiques du logement actuel, etc.

Si, par surcroît, on suppose que la relation entre les variables indépendantes et la variable latente est linéaire, on obtient le modèle suivant :

$$w_i = \sum_j x_{ij}\beta_j + u_i$$

où le terme aléatoire u_i représente les variations aléatoires entre les sujets (les ménages). Le modèle ne fait donc pas l'hypothèse que les sujets sont tous identiques, mais seulement que le modèle est suffisamment complet pour qu'on puisse considérer les variations de comportements non expliquées par le modèle comme l'effet du hasard.

Intégration des deux parties du modèle

On combine les deux parties du modèle.

$w_i < 0$ équivaut à $\sum_j x_{ij}\beta_j + u_i < 0$, c'est-à-dire à $u_i < -\sum_j x_{ij}\beta_j$

$w_i > 0$ équivaut à $\sum_j x_{ij}\beta_j + u_i \geq 0$, c'est-à-dire à $u_i \geq -\sum_j x_{ij}\beta_j$

Donc

$y_i = 0$ si $u_i < -\sum_j x_{ij}\beta_j$ et $y_i = 1$ si $u_i \geq -\sum_j x_{ij}\beta_j$

Ainsi, la probabilité que $y_i = 0$ est égale à la probabilité que $u_i < -\sum_j x_{ij}\beta_j$ et la probabilité que $y_i = 1$ est égale à la probabilité que $u_i \geq -\sum_j x_{ij}\beta_j$. Ces probabilités dépendent évidemment des hypothèses que l'on fait quant à la distribution des u_i. Selon l'hypothèse qu'on fait sur la fonction de densité de probabilité des u_i, on aura un modèle probit ou logit :

Hypothèse A : les u_i ont des distributions normales, indépendantes les unes des autres, avec moyenne nulle et variance commune σ^2 (hypothèses de Gauss-Markov) ; c'est le modèle probit.

Hypothèse B : les u_i sont distribués indépendamment les uns des autres et leurs fonctions de distribution cumulatives sont données par la fonction logistique ; c'est le modèle logit.

4-3.1.3 LE MODÈLE LOGIT ET L'INDUCTION STATISTIQUE

Avec un modèle comportant une variable dépendante latente (inobservable), comme celui qui précède, il est évidemment impossible d'estimer les paramètres β_j avec la méthode des moindres carrés. Il faut recourir à la méthode du maximum de vraisemblance. La première étape dans l'application de cette méthode est de construire la fonction de vraisemblance. Celle-ci, on se le rappelle, est la fonction de probabilité de l'échantillon exprimée en fonction des paramètres [2]. La fonction de vraisemblance découle donc de l'hypothèse qui est faite quant à la distribution des u_i, qui constitue une spécification complète du modèle aléatoire.

Nous nous attachons ici plus particulièrement au modèle logit, qui découle de l'hypothèse d'une distribution logistique. La fonction logistique est la fonction

$$L(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t + 1}$$

Si la variable aléatoire u a une fonction de distribution cumulative logistique, on a

$$\Pr[u < t] = L(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t + 1}$$

---

[2] Lorsque nous avons présenté le principe du maximum de vraisemblance (chapitre 2-2), nous avons défini la fonction de vraisemblance comme la fonction de densité de probabilité de l'échantillon. Ici toutefois, la variable observée est discrète (dichotomique, en l'occurrence) : la fonction de vraisemblance est donc la fonction de probabilité tout court.

Le choix de la fonction logistique peut se justifier par des arguments pragmatiques : elle est relativement commode à manipuler et elle est proche de la normale autour de la moyenne (elle s'en éloigne cependant pour les valeurs extrêmes). Toutefois, le choix de la fonction logistique peut s'appuyer sur des fondements théoriques beaucoup plus solides [3].

Les tests d'hypothèse que l'on peut faire sont basés sur des distributions *asymptotiques*. Ils ne sont donc valides que lorsque l'échantillon est suffisamment grand ; autrement dit, on aura d'autant plus confiance à ces tests que l'échantillon sera grand.

Les estimateurs b_j des paramètres β_j ont des distributions asymptotiques normales ; les estimateurs de leurs écarts type ont une distribution asymptotique χ^2. Ces conditions étant réunies, on peut appliquer les test *t* de Student aux paramètres.

4-3.2 Vers le logit multinomial : une généralisation heuristique du binomial

De toute évidence, le modèle stimulus-réaction, avec seuil de réaction, ne peut s'appliquer qu'à une situation de choix binaire. S'il y a plus de deux possibilités, il faut un modèle plus général, où le sujet choisit la possibilité qui présente pour lui le plus grand attrait.

Exemple : choix de la langue d'usage

- $y_i = 0$ si la langue d'usage est le français ;
- $y_i = 1$ si la langue d'usage est l'anglais ;
- $y_i = 2$ si la langue d'usage est l'italien ;
- $y_i = 3$ si la langue d'usage est une autre langue.

Le modèle théorique de comportement sous-jacent au modèle logit multinomial est le modèle d'utilité aléatoire (*random utility*). Selon ce modèle, il y a une variable latente qui mesure l'« utilité » ou l'attrait de chaque option, en fonction des attributs du sujet et des caractéristiques de l'option. Le sujet choisit rationnellement la possibilité qui a le plus grand attrait [4]. L'utilité de chaque option n'est cependant pas une fonction déterministe des variables indépendantes. Elle est constituée de la somme de deux termes : l'utilité *systématique* et un terme aléatoire.

---

3 Voir notamment Ben Akiva et Lerman (1985).

4 Le modèle d'utilité aléatoire se distingue en cela des modèles avec utilité « constante », où la probabilité qu'une option soit choisie croît avec son utilité, mais où il n'est pas certain qu'un sujet choisisse la possibilité la plus avantageuse pour lui. Dans les modèles avec utilité « constante », le choix de la possibilité la plus avantageuse est seulement plus probable que les autres. On s'écarte en cela de l'hypothèse de comportement rationnel.

Dans une situation de choix polytomique, au lieu d'une seule variable non observée (« latente ») w, qui mesure le stimulus, il faut autant de variables latentes qu'il y a de possibilités, chacune mesurant l'attrait de la possibilité correspondante. Écrivons

w_{ij} : mesure de l'attrait de la possibilité j pour l'individu i.

Dans un premier temps, reformulons le modèle binaire suivant ce schéma. Cela donne

$y_i = 0$ si $w_{i0} > w_{i1}$

$y_i = 1$ si $w_{i1} > w_{i0}$

ce qui est équivalent à

$y_i = 0$ si $w_{i0} - w_{i1} > 0$, c'est-à-dire si $w_{i1} - w_{i0} < 0$

$y_i = 1$ si $w_{i1} - w_{i0} > 0$

On peut donc interpréter la mesure du stimulus w_i du modèle binaire comme une mesure de la différence entre l'attrait de la première et de la seconde possibilité. Le modèle du stimulus devient

$w_i = w_{i1} - w_{i0} = z_{i1} - z_{i0} + u_i$

où $z_{ij} = \sum_h x_{ih}\beta_{hj}$ est la partie déterministe du modèle de l'attrait de la possibilité k.

Soit

$p_{ij} = \Pr[y_i = j]$: la probabilité que l'individu i choisisse la possibilité j.

On a

$$p_{i1} = \frac{e^{z_{i1} - z_{i0}}}{e^{z_{i1} - z_{i0}} + 1}$$

$$p_{i0} = 1 - p_{i1} = \frac{1}{e^{z_{i1} - z_{i0}} + 1}$$

ce qui implique, pour la cote (*odds*) de $[y_i = 1]$ contre $[y_i = 0]$:

$$\frac{p_{i1}}{1 - p_{i1}} = \frac{p_{i1}}{p_{i0}} = e^{z_{i1} - z_{i0}} = \frac{e^{z_{i1}}}{e^{z_{i0}}}$$

Supposons, sans le démontrer, que ce résultat puisse être généralisé à un nombre quelconque de possibilités et que, pour toute paire de possibilités j,k :

$$\frac{p_{ij}}{p_{ik}} = e^{z_{ij} - z_{ik}} = \frac{e^{z_{ij}}}{e^{z_{ik}}}$$

En faisant la somme sur toutes les possibilités j, on obtient

$$\sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

où $\sum_j p_{ij} = 1$, de sorte que

$$\sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j p_{ij}}{p_{ik}} = \frac{1}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

d'où

$$p_{ik} = \frac{e^{z_{ik}}}{\sum_j e^{z_{ij}}}$$

Et voilà le modèle logit multinomial !

On estime les paramètres de ce modèle au moyen de la méthode du maximum de vraisemblance, comme pour le logit binomial. Les procédures d'induction statistique sont analogues.