

MODÈLES À VARIABLE DÉPENDANTE QUALITATIVE

Problématique

Variable dépendante dichotomique :

$$y_i = 0 \text{ ou } 1$$

Exemple : mobilité résidentielle des ménages

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

La régression linéaire n'est pas applicable.

Un premier élément de solution

La véritable variable dépendante est la *probabilité* que $y = 1$:

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

MAIS

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.

MODÈLES À VARIABLE DÉPENDANTE QUALITATIVE

MODÈLE DE COMPORTEMENT

Modèle de comportement

- Variable non observée (« latente ») w : mesure le **stimulus** auquel est soumis le sujet
- Seuil de réaction : S

1. MODÈLE DE RÉACTION

$$y_i = 0 \text{ si } w_i < S$$

$$y_i = 1 \text{ si } w_i > S$$

Ajustement de l'échelle : w (variable ordinaire) définie de telle façon que $S = \text{zéro}$:

$$y_i = 0 \text{ si } w_i < 0$$

$$y_i = 1 \text{ si } w_i > 0$$

2. MODÈLE DU STIMULUS TOTAL

$$w_i = f(x_{i1}, x_{i2}, \dots, x_{ik})$$

$$\text{et si la relation est linéaire, } w_i = \sum_j x_{ij} \beta_j + u_i$$

- Le terme aléatoire u_i représente les variations aléatoires entre les sujets.

3. INTÉGRATION DES DEUX PARTIES DU MODÈLE

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

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

Donc

$$y_i = 0 \text{ si } u_i < -\sum_j x_{ij} \beta_j$$

$$y_i = 1 \text{ si } u_i > -\sum_j x_{ij} \beta_j$$

C'est-à-dire

$$\text{Prob}[y_i = 0] = \text{Prob}[u_i < -\sum_j x_{ij} \beta_j]$$

$$\text{Prob}[y_i = 1] = \text{Prob}[u_i > -\sum_j x_{ij} \beta_j]$$

MODÈLES À VARIABLE DÉPENDANTE QUALITATIVE

MODÈLE D'ÉCHANTILLONNAGE

Probit ou logit ?

Probit : les u_i ont des distributions normales, indépendantes les unes des autres, avec moyenne nulle et variance commune σ^2 (hypothèses de Gauss-Markov)

Logit : les u_i sont distribués indépendamment les uns des autres et leurs fonctions de distribution cumulatives sont données par la fonction logistique

La distribution logistique comme modèle d'échantillonnage

Fonction logistique

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

Distribution cumulative logistique

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

Si u_i a une distribution logistique :

$$\Pr[y_i = 0] = \Pr[u_i < -\sum_j x_{ij} \beta_j]$$

alors

$$\Pr[y_i = 0] = \Pr\left[u_i < -\sum_j x_{ij} \beta_j\right] = L\left(-\sum_j x_{ij} \beta_j\right) = \frac{1}{1 + e^{-\sum_j x_{ij} \beta_j}} = \frac{e^{\sum_j x_{ij} \beta_j}}{e^{\sum_j x_{ij} \beta_j} + 1}$$

⇒ Estimation par la méthode du maximum de vraisemblance

Le modèle logit et l'induction statistique

- Les tests d'hypothèse que l'on peut faire sont basés sur des distributions *asymptotiques*.
- 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.
- Tests d'hypothèse du rapport de vraisemblance r : $-2 \log r$ est asymptotiquement χ^2 .

LE MODÈLE LOGIT MULTINOMIAL

Variable dépendante polytomique

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.

Modèle théorique de comportement du logit multinomial

- Modèle théorique de comportement : modèle d'utilité aléatoire (*random utility*)
 - Il y a une variable latente mesure l'« utilité » ou l'attrait de chaque option,
 - en fonction des attributs du sujet et des caractéristiques de l'option.
 - w_{ij} : mesure de l'attrait de la possibilité j pour l'individu i .
 - Le sujet choisit rationnellement la possibilité qui a le plus grand attrait.
 - 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.

LE MODÈLE LOGIT MULTINOMIAL GÉNÉRALISATION HEURISTIQUE DU BINOMIAL (1)

Modèle binomial reformulé

MODÈLE DE RÉACTION

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

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

équivalent à

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

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

Le w_i du modèle binaire = différence entre l'attrait de la première et de la seconde possibilité.

MODÈLE DU STIMULUS

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

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

LE MODÈLE LOGIT MULTINOMIAL GÉNÉRALISATION HEURISTIQUE DU BINOMIAL (2)

Développement du modèle

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

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

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

La cote (*odds*) de $[y_i = 1]$ contre $[y_i = 0]$: $\frac{p_{i1}}{1 - p_{i1}} = \frac{p_{i1}}{p_{i0}} = e^{z_{i1} - z_{i0}} = \frac{e^{z_{i1}}}{e^{z_{i0}}}$

Généralisation heuristique à un nombre quelconque de possibilités :

$$\text{pour toute paire de possibilités } j, k : \frac{p_{ij}}{p_{ik}} = e^{z_{ij} - z_{ik}} = \frac{e^{z_{ij}}}{e^{z_{ik}}}$$

$$\text{Somme sur toutes les possibilités } j : \sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

$$\text{où } \sum_j p_{ij} = 1, \text{ de sorte que } \sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j p_{ij}}{p_{ik}} = \frac{1}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

$$\text{d'où } p_{ik} = \frac{e^{z_{ik}}}{\sum_j e^{z_{ij}}}$$