URCA Hugo Harari-Kermadec

2008-2009 harari@ecogest.ens-cachan.fr

Économétrie 2 : probit multivarié

Ce cours est fortement inspiré de celui de Christophe Hurlin,
`http://www.univ-orleans.fr/deg/masters/ESA/CH/Qualitatif_Chapitre2.pdf` .

I Introduction

Les modèles Probit et Logit simples que nous avons vu dans le chapitre précédent sont adaptés au cas d'un variable qualitative dichotomique. Lorsque la variable à expliquer est discrète mais prend plus de deux valeurs, il faut adapter la méthode. On sépare en fait plusieurs cas :

Modèles multinomiaux ordonnés
Les différentes valeurs de Y représentent des classes de valeurs ordonnées, par exemple des tranches de revenus ou des classes d'age.

Modèles multinomiaux séquentiels
Les valeur de Y indiquent la survenue d'événement consécutif, par exemple porteur sain d'une maladie, puis malade puis décédé.

Modèles multinomiaux non ordonnés
Les données correspondent à des qualités différentes, comme les catégories socio-professionnelles ou le choix d'un candidat.

Dans le cadre le plus général, on peut supposer que le nombre de modalité pour Y dépend de l'individu. Nous nous restreindrons au cas plus simple où on suppose ce nombre fixe pour tout l'échantillon. On se donne une valeur de référence, que l'on note 0, et on le note numérote les autres valeurs possibles de 1 à m. Il y a donc $m+1$ valeurs possibles avec $m > 1$ (sinon on est dans le cas dichotomique).
Le problème est alors résumé par la recherche des m probabilités $\mathbb{P}(Y=1)$, $\mathbb{P}(Y=2)$, ...et $\mathbb{P}(Y=m)$ puisque $\mathbb{P}(Y=0) = 1 - \sum_{j=1}^{m} \mathbb{P}(Y=j)$. Chacune de ces probabilités s'écrit comme une fonction des variables explicatives X_i et du vecteur des paramètres θ :

$$\mathbb{P}(Y_i = j) = G_j(X_i, \theta).$$

Soit $y_{ij} = \mathbb{1}_{Y_i=j}$. La vraisemblance s'écrit alors

$$L(\theta) = \prod_{j=0}^{m} \left[\prod_{i:\ Y_i=j} \mathbb{P}(Y_i = j) \right] = \prod_{i=1}^{n} \prod_{j=0}^{m} \mathbb{P}(Y_i = j)^{y_{ij}} = \prod_{i=1}^{n} \prod_{j=0}^{m} G_j(X_i, \theta)^{y_{ij}}.$$

On peut alors obtenir un estimateur de maximum de vraisemblance une fois les fonctions G_j convenablement choisies. Dans la suite, nous allons présenter les fonctions adaptées aux trois types de modèles multinomiaux.

II Modèles multinomiaux ordonnés

II.1 Modélisation

Dans le cas des modèles multinomiaux ordonnés, Y peut naturellement s'écrire à l'aide d'une variable latente comme une généralisation du modèle Logit simple :

$$Y_i = \begin{cases} 0 & \text{si } Y_i^* \leq c_1 \\ 1 & \text{si } c_1 < Y_i^* \leq c_2 \\ \dots & \dots \\ m & \text{si } c_m < Y_i^*. \end{cases}$$

où les c_j sont en ordre croissants et où la variable latente Y_i^* suit un modèle linéaire :

$$Y_i^* = X_i\theta + \varepsilon_i.$$

Attention : dans cette paramétrisation, le niveau de Y^* est paramétré par les c_j, on n'introduit donc pas de constante dans le modèle linéaire (sinon, le modèle ne serait pas identifiable) :

$$\theta = (\theta_1, \ldots, \theta_K)' \qquad \text{(pas de } \theta_0\text{)}.$$

Comme dans le cas dichotomique, ε_i est supposé d'espérance nulle et de variance 1, de loi gaussienne ou logistique. On a alors les modèles Probit multinomial ordonné et Logit multinomial ordonné. Soit F la fonction de répartition de la loi choisie, on a alors :

$$\begin{array}{rcccl} \mathbb{P}(Y_i = 0) & = & \mathbb{P}(Y_i^* \leq c_1) & = & F(c_1 - X_i\theta) \\ \mathbb{P}(Y_i = 1) & = & \mathbb{P}(c_1 < Y_i^* \leq c_2) & = & F(c_2 - X_i\theta) - F(c_1 - X_i\theta) \\ & & \ldots & & \\ \mathbb{P}(Y_i = m) & = & \mathbb{P}(c_m < Y_i^*) & = & 1 - F(c_m - X_i\theta) \end{array}$$

En posant $c_0 = -\infty$ et $c_{m+1} = +\infty$, on a :

$$\mathbb{P}(Y_i = j) = \mathbb{P}(c_j < Y_i^* \leq c_{j+1}) = F(c_{j+1} - X_i\theta) - F(c_j - X_i\theta).$$

La vraisemblance s'écrit donc

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=0}^{m} \left[F(c_{j+1} - X_i\theta) - F(c_j - X_i\theta)\right]^{y_{ij}}.$$

II.2 Tests

Comme dans le logit simple, le test le plus naturel consiste à construire un rapport de vraisemblance. Pour tester une contrainte de rang r sur θ de dimension p, on utilise le résultat suivant :

$$LR = -2\log L(\hat{\theta}) - \log L(\hat{\theta}^c) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_r,$$

où $\hat{\theta}^c$ est l'estimateur du maximum de vraisemblance sous la contrainte.

II.3 Implémentation sous SAS

Les mots en majuscule sont des commandes SAS. Les mots en minuscule sont des noms donnés par l'utilisateur. La procédure est la même que pour le logit simple, les sorties sont donc pratiquement identiques. On indique ici les principales différences.

II.3.1 Entrée

Il faut à nouveau utiliser l'option DESC puisqu'on veut bien comparer à la valeur de référence 0. On tape par exemple :
PROC LOGISTIC DATA=donnees DESC ;
MODEL Y = X1 X2 / LINK= LOGIT ;
RUN ;

II.3.2 Sortie

La sortie est très proche. On a par contre plus de constantes (INTERCEPT) que dans le modèle simple : il y a m constantes (les c_j) si Y prend $m+1$ valeurs (il n'y a pas de seuil pour la valeur 0).

III Modèles multinomiaux séquentiels

III.1 Modélisation

Les événements s'enchaînant, on parle souvent d'étapes. Dans ce type de modèle, se sont les probabilités conditionnelles qui ont un sens pertinent. On peut voir ce modèle comme une suite de Logit simple emboîté : réussite ou échec à la première étape ; puis, conditionnellement à la réussite à la première étape, réussite ou échec à la seconde... On suppose l'existence de variables latentes indiquant la réussite à l'étape t conditionnellement au passage de l'étape $t-1$:

$$Y_{it}^* = X_i\theta + \varepsilon_{it}$$

où ε_{it} suit une loi gaussienne ou logistique, de fonction de répartition F_t. La probabilité conditionnelle d'arrêt à l'étape t est donc $P(Y_{it} \leq 0)$, c'est-à-dire $F_t(X_i\theta)$. "$Y_i = j$" signifie que l'on s'arrête à l'étape j, c'est-à-dire qu'on passe chacune des étapes précédentes mais pas la j-ième. On a donc

$$\mathbb{P}(Y_i = j) = F_j(X_i\theta) \prod_{t=1}^{j-1} (1 - F_t(X_i\theta))$$

Le cas $j = 0$ est donné par l'échec à l'étape 1 : $\mathbb{P}(Y_i = 0) = F_1(X_i\theta)$.

III.2 Implémentation

Cette méthode n'est pas implémentée sous SAS.

IV Modèles multinomiaux non ordonnés

IV.1 Modélisation

Cette fois-ci, les différentes valeurs de Y correspondent à des choix non ordonnés. Une hypothèse essentielle est nécessaire ici, l'hypothèse d'Indépendence des Alternatives Non Pertinentes (IANP) : on suppose que l'existence du choix C n'influe pas sur le rapport des probabilités de A et B. Cette hypothèse est très forte et rarement justifiable sur des exemples concrets. Elle est néanmoins admise la plupart du temps.
On peut prendre un exemple pour expliquer cette hypothèse : dans le cadre de l'élection présidentielle de 2007, l'IANP implique que le choix entre Bayrou et Sarkozy ne dépend pas de l'existence de la candidate Royal. L'hypothèse n'est pas vérifiée, puisqu'en l'absence de Royal, les sondages donnaient Bayrou vainqueur. Par contre, en présence de Royal, et d'autres candidats, c'est Sarkozy qui est sorti vainqueur de l'élection.
Le modèle est différent des précédents. On utilise la loi logistique mais cette fois-ci, le vecteur des paramètres dépend du choix $j \in [|0; m|]$:

$$\mathbb{P}(Y_i = j) = \frac{\exp(X_i\theta^{[j]})}{1 + \sum_{j=1}^{m} \exp(X_i\theta^{[j]})}.$$

L'identifiabilité est donnée par la convention $\theta^{[0]} = 0 \in \mathbb{R}^K$ et la somme des probabilité est égale à 1.
Comme précédemment, il n'y a pas de constante (elle est comprise dans les seuils c_j), c'est pourquoi les vecteurs de paramètres $\theta^{[j]}$ sont dans $\mathbb{R}^K$.
Le rapport entre les probabilités de choix est donc donné par :

$$\frac{\mathbb{P}(Y_i = j)}{\mathbb{P}(Y_i = l)} = \frac{\exp(X_i\theta^{[j]})}{\exp(X_i\theta^{[l]})} = \exp\left(X_i(\theta^{[j]} - \theta^{[l]})\right).$$

On peut alors lire l'effet d'une variable explicative, par exemple la première, sur le choix de j plutôt que l : il est donné par le signe de $\theta_1^{[j]} - \theta_1^{[l]}$.

IV.2 Implémentation sous SAS

La procédure est la même que pour le modèle multinomial ordonné, la différence est dans le LINK : on utilise un modèle logistique généralisé.

IV.2.1 Entrée

La différence dans l'option LINK :
PROC LOGISTIC DATA=donnees DESC ;
MODEL Y = X1 X2 / LINK= GLOGIT ;
RUN ;

IV.2.2 Sortie

La sortie est très proche. On a encore une fois plus de constantes (INTERCEPT) que dans le modèle simple : il y a m constantes (les c_j) si Y prend $m + 1$ valeurs (il n'y a pas de seuil pour la valeur 0). Mais il y a aussi plusieurs valeurs pour les effets de chaque variable explicative : on estime l'effet de chaque variable explicative k sur l'augmentation du rapport des probabilités entre chaque choix et la référence : $\theta_k^{[j]}$.

V TP SAS

On va poursuivre l'étude du jeu de donnée sur le télé-achat. Toujours en tapant harari et reims dans google, vous devriez trouver ma page web vous concernant. Sinon, l'adresse est
`http://www.crest.fr/ckfinder/userfiles/files/Pageperso/hharari/harari_fichiers/reims.htm`
Il faut expliquer si la vente est mauvaise (0), moyenne (1) ou très bonne (2) en fonction des variables explicatives : jour de semaine ou de week-end, temps d'exposition du produit, niveau de réduction, émission en direct ou non.