Application : Modèles à choix binaires

Estimation des modèles Probit et Logit binaires explicatifs des facteurs de la réussite en Licence

Nous avons relevé sur un échantillon de 60 étudiants inscrits en dernière année de Licence d'Économie, les variables suivantes susceptibles d'expliquer la réussite ou l'échec à l'examen de Licence (variable REUSSITE = 0 si échec, 1 sinon) :

NENFANTS = variable discrète représentant le nombre de frères et soeurs de l'étudiant,

NECONO = la note d'économétrie sur 20 obtenue en Licence,

NMICRO = la note de micro-économie sur 20 obtenue en Licence,

GENRE = variable muette, (1 = masculin, 0 = féminin).

Un extrait des données est présenté dans le tableau 1.

Tableau 1 – Extrait de données

OBS	REUSSITE	NENFANTS	NECONO	NMICRO	GENRE
1	0	2	3,6	0	1
2	0	5	3,8	0	1
...	...	...	...	...	...
59	1	0	16,2	12	0
60	1	2	17	4	0

$y^* = a_0 + a_1 Nenf + a_2 NEco + a_3 NMicro + a_4 Genre + \varepsilon_i$

On demande :

1) d'estimer un modèle de type Logit permettant de prévoir la probabilité de réussite d'un étudiant en Licence,

$\varepsilon_i \sim$ fonct logistique

2) de comparer les résultats avec un modèle de type Probit

3) de donner la probabilité de réussite, à l'aide du modèle Logit estimé, pour un étudiant dont les caractéristiques sont les suivantes : NENFANTS = 1 ; NECONO = 12 ; NMICRO = 13,5 ; GENRE = masculin.

Solution

1) Une première estimation d'un modèle Logit, conduit aux résultats suivants :

Dependent Variable : REUSSITE
Method: ML – Binary Logit
Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.682523	0.378870	– 1.801470	0.0716
NECONO	0.632062	0.239564	2.638382	0.0083
GENRE	– 3.761718	1.437068	– 2.617633	0.0089
NMICRO	0.155322	0.188916	0.822173	0.4110
C	– 3.265634	2.020060	–1.616602	0.1060

À la lecture des résultats, nous constatons que :

– la variable NMICRO à une probabilité critique de 0,41, elle n'est donc pas significative,

– la variable NENFANTS à une probabilité critique de 0,07, elle est donc faiblement significative.

Nous procédons à une nouvelle estimation en retirant la variable NMICRO dont le coefficient n'est pas significativement différent de 0.

Les résultats complets fournis par Eviews sont les suivants :

Dependent Variable : REUSSITE
Method: ML – Binary Logit
Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.746742	0.378942	– 1.970596	0.0488
NECONO	0.695857	0.231789	3.002112	0.0027
GENRE	– 3.634605	1.410945	– 2.576008	0.0100
C	– 2.859277	1.910377	– 1.496708	0.1345

Mean dependent var	0.516667	S.D. dependent var	0.503939
S.E. of regression	0.287086	Akaike info criterion	0.645890
Sum squared resid	4.615432	Schwarz criterion	0.785512
Log likelihood	– 15.37669	Hannan-Quinn criter.	0.700504
Restr. log likelihood	– 41.55549	Avg. log likelihood	– 0.25627
LR statistic (3 df)	52.35761	McFadden R-squared	0.629972
Probability(LR stat)	2.51E – 11		

Obs with Dep=0	29	Total obs	60
Obs with Dep=1	31		

Avec : $L_U =$ Log likelihood ; $L_R =$ Restr. log likelihood ; $LR = LR$ statistic ; $L_U/n =$ Avg. log likelihood. Le critère d'information de Hannan-Quinn permet des comparaisons entre les modèles (comme les critères de Akaike ou Schwarz) en termes d'arbitrage : apport d'information lié à l'ajout de variables explicatives et perte de degrés de liberté. En cas de modèle concurrent, celui ayant le plus faible critère d'information sera retenu.

a) Interprétation statistique

Les coefficients sont tous significativement différents de 0, hormis le terme constant.

La statistique de la Log vraisemblance est égale à $LR = 52,35$ que l'on compare à un χ^2 lu dans la table à un seuil de 0,95 % et à 3 degrés de liberté. $\chi_3^{2\,0,95} = 9,28 < 52,35$ $\longrightarrow$ rejet de H0.

Le pseudo-R^2 est donné par :

$$R^2 = 1 - \frac{Log(Lu)}{Log(L_R)} = 1 - \frac{-15,38}{-41,56} = 1 - 0,37 = 0,63$$

Le modèle est validé sur le plan statistique.

Le Logiciel permet d'élaborer la table de succès de prédiction suivante :

		0	Prédit ($\hat{y}_i$)	1	Total
Observé (y_i)	0	26		4	30
	1	3		27	30
Total		29		31	60

On peut constater que :

- la proportion des prédictions correctes est égale à : $\frac{26+27}{60} = 88,33\%$.

- le pourcentage des prédictions fausses est égal à : $\frac{3+4}{60} = 11,67\%$.

Le taux d'erreur est donc faible indiquant une bonne qualité prédictive du modèle.

b) Interprétation économique

Le modèle s'écrit :

$$\text{Ln}\left(\frac{P_i}{1-P_i}\right) = \underset{(1,97)}{-0,75}\,\text{NENFANTS} + \underset{(3,00)}{0,70}\,\text{NECONO} \underset{(2,57)}{-3,63}\,\text{GENRE} - 2,86 + e_i$$

(.) = z-Statistique

e_i = Résidu d'estimation

– Le nombre de frères et soeurs du foyer agit négativement, les étudiants issus de familles nombreuses ont un taux de réussite plus faible.

– La note d'économétrie est un facteur positif de réussite.

– Enfin, les étudiants de genre masculin réussissent en général moins bien (signe négatif) que les étudiants de genre féminin.

2) Estimation d'un modèle Probit

L'estimation d'un modèle Probit conduit aux résultats suivants :

Dependent Variable : REUSSITE
Method: ML – Binary Probit
Included observations: 60

Variable	Coefficient	Std. Error	z-Statistic	Prob.
NENFANTS	– 0.428197	0.219223	– 1.953247	0.0508
NECONO	0.363148	0.110230	3.294454	0.0010
GENRE	– 1.824203	0.650426	– 2.804629	0.0050
C	– 1.491466	1.108767	– 1.345157	0.1786

Les valeurs des coefficients sont de même signe mais différentes par rapport au modèle Logit car la spécification n'est pas la même. Cependant, nous pouvons retrouver, approximativement, les valeurs estimées du modèle Logit en multipliant chacun des coefficients des variables explicatives par la constante[1] $\pi\sqrt{3} \approx 1{,}81$.

3) Soit les caractéristiques de l'étudiant : NENFANTS = 1 ; NECONO = 12 ; NMICRO = 13,5 ; GENRE = masculin. $\longrightarrow$ $\hat{P}_i = P(y_i = 1)$ Remb = 1

Le modèle Logit estimé (la note de micro-économie ne figurant pas dans le modèle final, elle n'est pas intégrée dans le calcul, *cf.* question 1) est le suivant :

$$\text{Ln}\left(\frac{P_i}{1-P_i}\right) = \underset{(1.97)}{-0{,}75}\,\text{NENFANTS} + \underset{(3.00)}{0{,}70}\,\text{NECONO} - \underset{(2.57)}{3{,}63}\,\text{GENRE} - 2{,}86 + e_i$$

$$\text{Ln}\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = -0{,}75 \times 1 + 0{,}70 \times 12 - 3{,}63 \times 1 - 2{,}86 = 1{,}109$$

$$\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = e^{1.109} = 3{,}033 \longrightarrow \hat{P}_i = 3{,}033/(1+3{,}033) = 0{,}75$$

La probabilité de réussite de cet étudiant de licence est donc de 75 %.

Application : Logit ordonné

Estimation d'un modèle à choix multiples de prévision des ventes

La société Télé-Ventes (ventes par téléphone lors d'une émission à la télévision) désire estimer le niveau des ventes par article pour chaque émission afin de dimensionner la charge de l'entrepôt et prévoir ainsi le nombre d'équipes.

L'émission est diffusée tous les jours sauf le dimanche. Les ventes sont réparties en trois classes : faible, moyenne, forte.

L'objectif est d'estimer un modèle permettant de prévoir à quelle classe de vente (faible, moyenne, forte) appartient un article présenté lors d'une émission. Pour ce faire, on dispose des informations suivantes sur 82 émissions passées :

VENTES : classe de l'article (faible = 0, moyenne = 1, forte = 2),

WE : variable indicatrice du type de jour de diffusion de l'émission (1 les jours de semaine, 0 le samedi),

EXPO : temps d'exposition du produit en minutes,

REDUC : % de réduction proposé sur le prix,

DIRECT = variable indicatrice d'émission enregistrée (0 pas direct, 1 direct).

Un extrait des données est présenté dans le tableau suivant:

Extrait de données

Obs	VENTES	WE	EXPO	REDUC	DIRECT
1	0	1	3,5	0,20	0
2	1	0	3,5	0,20	0
...	...	...	...	...	...
81	0	1	7	0	0
82	0	1	3,5	0	0

1) On demande d'estimer un modèle Logit multinomial permettant de prévoir la classe de vente d'un article à partir des facteurs explicatifs proposés.

2) D'effectuer une prévision pour un article présenté lors d'une émission en différé diffusée en semaine dont le temps d'exposition est de 7 minutes et sans réduction.

Solution

1) Une première estimation à l'aide d'un modèle de type Logit conduit au résultat suivant :

Dependent Variable : VENTES
Method: ML – Ordered Logit (Quadratic hill climbing)
Included observations: 82 after adjusting endpoints

	Coefficient	Std. Error	z-Statistic	Prob.
DIRECT	0.226809	1.325269	0.171142	0.8641
EXPO	0.644804	0.150121	4.295228	0.0000
REDUC	8.922215	2.759073	3.233772	0.0012
WE	– 1.395159	0.512093	– 2.724426	0.0064

Nous constatons que la variable DIRECT dont le coefficient est affecté d'une probabilité critique de 0,86 n'est pas significative (les téléspectateurs ne sont pas sensibles aux émissions diffusées en direct), elle est donc retirée du modèle. La nouvelle estimation est alors la suivante.

Dependent Variable : VENTES
Method: ML – Ordered Logit (Quadratic hill climbing)
Included observations: 82 after adjusting endpoints

	Coefficient	Std. Error	z-Statistic	Prob.
EXPO	0.641023	0.148189	4.325700	0.0000
REDUC	8.982926	2.735164	3.284237	0.0010
WE	– 1.377785	0.500824	– 2.751038	0.0059
		Limit Points		
LIMIT_1:C(4)	3.679581	0.994565	3.699689	0.0002
LIMIT_2:C(5)	6.170926	1.174856	5.252497	0.0000
Akaike info criterion	1.605306	Schwarz criterion		1.752057
Log likelihood	– 60.81754	Hannan-Quinn criter.		1.664224
Restr. log likelihood	– 79.48219	Avg. log likelihood		– 0.74167
LR statistic (3 df)	37.32930	LR index (Pseudo-R2)		0.234828
Probability(LR stat)	3.92E-08			

a) Interprétation statistique

Les coefficients sont maintenant tous significativement différents de 0 (probabilités critiques inférieures à 0,05).

La statistique de la Log vraisemblance est égale à $LR = 37{,}33$ que l'on compare à un χ^2 lu dans la table à un seuil de 0,95 % et à 3 degrés de liberté, $\chi_3^{2(0,95)} < 37{,}33 \longrightarrow$ rejet de H0.

Le modèle est donc validé sur le plan statistique.

b) Interprétation économique

La durée d'exposition et le pourcentage de réduction agissent positivement sur les ventes.

Une émission diffusée un jour de semaine engendre moins de ventes qu'une émission diffusée le samedi.

Les seuils c_1 et c_2 sont respectivement de 3,679 et 6,170.

Les signes des coefficients sont conformes à l'intuition économique.

vente = 0 si $y^* \leq 3{,}67$

vente = 1 si $y^* \in]3{,}67 ; 6{,}A)$

vente = 2 si $y^* > 6{,}A$

Le logiciel Eviews propose une table permettant d'appréhender les qualités prévisionnelles du modèle sur l'échantillon :

Dependent Variable : VENTES
Method: ML – Ordered Logit (Quadratic hill climbing)
Included observations: 82 after adjusting endpoints
Prediction table for ordored dependent variable

Value y_i	Count	Count of obs with Max Prob	Error	Sum of all Probabilities	Error
0	44	47	−3	44.381	−0.381
1	27	31	−4	26.834	0.166
2	11	4	7	10.785	0.215

À la lecture des résultats, nous constatons que les qualités prévisionnelles de ce modèle sont satisfaisantes car le taux d'erreur est assez faible pour les ventes de niveau 0 et 1. En revanche, pour les ventes de niveau 2 (forte) nous constatons que sur 11 ventes réalisées, le modèle en a prévues correctement seulement 4.

2) Nous calculons l'estimation de la variable latente correspondante aux caractéristiques de l'émission avec EXPO = 7, REDUC = 0 et WE = 1 :

$$\widehat{y^*} = 0{,}641 \times \text{EXPO} + 8{,}982 \times \text{REDUC} - 1{,}377 \times \text{WE}$$

$$\widehat{y^*} = 0{,}641 \times 7 + 8{,}982 \times 0 - 1{,}377 \times 1 = 3{,}109$$

Puis nous calculons les probabilités :

$$P_1 = \text{Prob}(y_i = 0) = \Phi(c_1 - x_i\hat{a}) = \Phi(3{,}679 - 3{,}109)$$

$$= \Phi(0{,}57) = \frac{e^{0{,}57}}{1 + e^{0{,}57}} = 0{,}639$$

$$P_2 = \text{Prob}(y_i = 1) = \Phi(c_2 - x_i\hat{a}) - \Phi(c_1 - x_i\hat{a}) = \Phi(6{,}17 - 3{,}109) - 0{,}639$$

$$P_2 = \Phi(3{,}06) - 0{,}639 = \frac{e^{3{,}06}}{1 + e^{3{,}06}} - 0{,}639 = 0{,}955 - 0{,}639 = 0{,}31$$

$$P_3 = \text{Prob}(y_i = 2) = 1 - \Phi(6{,}17 - 3{,}109) = 1 - 0{,}955 = 0{,}045$$

Soit les résultats suivants :

	VENTES = 0	VENTES = 1	VENTES = 2
P_i	0,639	0,316	0,045

Les ventes de l'émission prévue ont 64% de probabilité d'appartenir à la classe 0 de faible vente.

Application : Le modèle Tobit

Prévision de la demande d'électricité pour un fournisseur à capacité limitée

Dans le cadre de l'ouverture du marché de l'électricité à destination des industriels, un fournisseur d'électricité, qui n'est pas l'opérateur historique, propose de l'énergie électrique à bas prix dans la limite de ses capacités fixées à 3 000 mégawatts : la demande supérieure à ce seuil ne peut donc pas être servie.

Au-delà de ses capacités les clients sont délestés et sont donc dans l'obligation de basculer vers une autre source. La demande (y_t) exprimée en mégawatts à la période t est fonction de trois facteurs explicatifs :

x_{1t} = indicateur d'écart de prix par rapport à la concurrence en t, la valeur indique le % de réduction, pour le jour considéré, accordé par l'opérateur historique (si 0 pas de réduction de prix),

x_{2t} = nombre de clients industriels alimentés en t,

x_{3t} = variable indicatrice signalant les jours particuliers à forte consommation tels que le lendemain de jour férié, ...

Soit les données quotidiennes sur 60 jours (*cf.* un extrait dans le tableau : dont cet opérateur dispose.

Tableau . · Extrait de données

Jour	y	x_1	x_2	x_3
1	2717	0	61	0
2	2126	0	32	0
...	...	...	...	...
59	2683	0,1	61	0
60	3000	0	79	0
61		1	98	1
62		0	60	0

On demande :

1) d'estimer un modèle Logit permettant de prévoir la demande quotidienne à partir des facteurs explicatifs proposés et de commenter les résultats ;

2) d'effectuer une prévision pour les jours 61 et 62 sachant que :
$x_{1,61} = 1$; $x_{2,61} = 98$; $x_{3,61} = 1$ et $x_{1,62} = 0$; $x_{2,62} = 60$; $x_{3,62} = 0$.

Solution

1) La consommation est censurée car les valeurs de la variable à expliquer (la demande) ne sont pas connues lorsqu'elles sortent de l'intervalle [0 ; 3 000] puisque au-delà de 3 000, la demande ne peut être satisfaite.

Les résultats d'estimation sont les suivants :

Dependent Variable : Y
Method: ML – Censored Normal (TOBIT) (Quadratic hill climbing)
Included observations: 60 after adjustments
Left censoring (value) series: 0
Right censoring (value) series: 3000

	Coefficient	Std. Error	z-Statistic	Prob.
X1	– 27.69316	7.582649	– 3.652175	0.0003
X2	20.50361	0.480937	42.63261	0.0000
X3	186.2357	34.88046	5.339256	0.0000
C	1473.642	23.60984	62.41642	0.0000
Error Distribution				
SCALE: SIG	47.41983	4.617038	10.27062	0.0000

R-squared	0.981952	Mean dependant var	2542.667
S.E. of regression	46.09561	Akaike info criterion	9.355239
Sum squared resid	116864.3	Schwarz criterion	9.529768
Log likelihood	– 275.6572	Hannan-Quinn criter	9.423507
Avg. log likelihood	– 4.594286		
Left censored obs	0	Right censored obs	8
Uncensored obs	52	Total obs	60

a) Interprétation statistique

Les coefficients sont tous significativement différents de 0 (les probabilités critiques des coefficients sont toutes inférieures à 0,05), le modèle est validé sur le plan statistique.

Eviews indique, sur l'avant dernière ligne, le nombre de données censurées : 0 à gauche et 8 à droite.

b) Interprétation économique

– L'indicateur d'écart de prix agit négativement sur la demande : en cas de réduction tarifaire de la concurrence, la demande diminue.

– Le nombre de clients connectés au réseau et la variable muette « type de jour » ont un effet positif sur la demande.

– La variable d'échelle (estimateur de σ) est égale à 47,41.

Les coefficients ont bien le signe attendu, le modèle est validé sur le plan économique.

Le modèle Tobit s'écrit d'après [6] :

$$\hat{y}_i^* = -27{,}69 \times x_{1i} + 20{,}50 \times x_{2i} + 186{,}23 \times x_{3i} + 1473{,}64$$

2) Le calcul de la prévision pour les jours 61 et 62 est directement effectué par application du modèle Tobit estimé.

Sachant que :

$$x_{1,61} = 1 \; ; x_{2,61} = 98 \; ; x_{3,61} = 1 \text{ et } x_{1,62} = 0 \; ; x_{2,62} = 60 \; ; x_{3,62} = 0.$$

La prévision pour le jour 61 est donnée par :

$$y^*_{61} = -27{,}69 \times 1 + 20{,}50 \times 98 + 186{,}23 \times 1 + 1473{,}64 = 3641{,}53$$

$$y^*_{61} > c_2 \longrightarrow y_{61} = c_2 = 3000$$

La prévision pour le jour 62 est donnée par :

$$y^*_{62} = -27{,}69 \times 0 + 20{,}50 \times 60 + 186{,}23 \times 0 + 1473{,}64 = 2703{,}85$$

$$c_1 < y^*_{62} < c_2 \longrightarrow y_{62} = y^*_{62} = 2703{,}85$$