

REGRESSION LOGISTIQUE (LOGIT)

Source : D.N.Gujarati, "Basic Econometrics", Third Ed., McGraw Hill, 1995

Le concept

Dans les modèles de régression classiques, la variable dépendante est une variable quantitative. Les modèles qui utilisent comme variable dépendante une variable binaire sont appelés modèles probabilistique linéaire. Ils sont la forme suivante :

$$Y_i = \mathbf{b}_1 + \mathbf{b}_2 X_i + u_i$$

avec $Y = 1$ pour une catégorie et

$Y = 2$ pour l'autre catégorie.

$E(Y_i|X_i)$ peut en effet être interprété comme la probabilité qu'un événement (représenté par Y_i) se produise conditionnellement à X_i . Si P_i est la probabilité que Y_i soit égal à 1 (et donc $1-P_i$, la probabilité que Y_i soit égal à 0), on a donc bien :

$$E(Y_i|X_i) = \mathbf{b}_1 + \mathbf{b}_2 X_i = P_i$$

Quels sont les problèmes que soulève l'utilisation de tels modèles ?

- les termes d'erreurs ne suivent pas une distribution gaussienne. En effet, $u_i = Y_i - \mathbf{b}_1 - \mathbf{b}_2 X_i$. Si $Y_i = 0$, $u_i = -\mathbf{b}_1 - \mathbf{b}_2 X_i$ et si $Y_i = 1$, $u_i = 1 - \mathbf{b}_1 - \mathbf{b}_2 X_i$. Les termes d'erreurs ne suivent donc pas une distribution gaussienne mais bien une distribution binomiale.
- hétéroscédasticité de la variance : sachant que $\text{var}(u_i) = E[u_i - E(u_i)]^2$ et que $E(u_i) = 0$, $\text{var}(u_i) = E(u_i^2)$. Les u_i sont égaux à $-\mathbf{b}_1 - \mathbf{b}_2 X_i$ avec probabilité $(1-P_i)$ et $1 - \mathbf{b}_1 - \mathbf{b}_2 X_i$ avec probabilité P_i . On peut en déduire que la variance des u_i est égale à $(\mathbf{b}_1 + \mathbf{b}_2 X_i)(1 - \mathbf{b}_1 - \mathbf{b}_2 X_i)$, soit $E(Y_i|X_i)[1 - E(Y_i|X_i)]$. La variance des résidus est donc conditionnelle aux X_i .
- $E(Y_i|X_i)$ est la probabilité d'observer un événement. Si, par construction Y_i est bien borné entre 0 et 1, rien ne garantit pour \hat{Y}_i . Il s'agit en effet du résultat d'une équation linéaire qui peut très bien prendre des valeurs en-dehors de l'intervalle attendu. Le résultat obtenu ne peut alors plus s'interpréter aisément comme une probabilité.
- enfin la mesure conventionnellement du R^2 perd une grande partie de son intérêt dans de tels modèles. La variable dépendante est en effet binaire alors que le résultat du modèle de régression est quantitatif. La mesure du carré du coefficient de corrélation entre une variable qualitative et une variable quantitative n'a guère de sens ...

Le modèle

Le modèle logistique offre des solutions à ces différents problèmes. Il prend la forme suivante :

$$P_i = E(Y = 1|X_i) = \frac{1}{1 + e^{-(\mathbf{b}_1 + \mathbf{b}_2 X_i)}}$$

Il s'agit d'une fonction logistique. Un tel modèle assure que les \hat{Y}_i sont bornés entre 0 et 1 mais il prend appui sur une relation non-linéaire entre les P_i et les X_i . Il ne peut donc être estimé par la procédure des moindres carrés. Une solution existe toutefois et est la suivante :

$$1 - P_i = \frac{1}{1 + e^{(b_1 + b_2 X_i)}}$$

$$\Rightarrow \frac{P_i}{1 - P_i} = \frac{1 + e^{(b_1 + b_2 X_i)}}{1 + e^{-(b_1 + b_2 X_i)}} = e^{(b_1 + b_2 X_i)}$$

$$\Rightarrow \ln\left(\frac{P_i}{1 - P_i}\right) = L_i = b_1 + b_2 X_i$$

Le rapport des probabilités est donc linéaire aussi bien en X que dans les paramètres du modèle.

L'interprétation du modèle est la suivante :

- b_2 mesure le changement du rapport entre les probabilités lorsque la variable explicative auquel il est associé augmente d'une unité;
- connaissant b_1 et b_2 , on peut estimer P_i directement par $P_i = E(Y = 1 | X_i) = \frac{1}{1 + e^{-(b_1 + b_2 X_i)}}$;

L'estimation du modèle

Pour procéder à l'estimation du modèle, il faut calculer les L_i , ce qui pose évidemment problème puisque l'on obtient :

- si $P_i = 1$: $L_i = \ln\left(\frac{1}{0}\right)$
- si $P_i = 0$: $L_i = \ln\left(\frac{0}{1}\right)$

Deux approches sont possibles pour résoudre ce problème. La première passe par les estimateurs du maximum de vraisemblance. La seconde, que nous utiliserons ici, suppose que l'on connaisse les fréquences relatives de l'événement étudié. Si le nombre d'observations effectués est suffisant, les fréquences relatives sont en effet une bonne approximation des probabilités théoriques¹. En utilisant des données groupées, du type de celles présentées ci-dessous, on peut donc procéder à une estimation du modèle par :

$$\hat{L}_i = \ln\left(\frac{\hat{P}_i}{1 - \hat{P}_i}\right) = \hat{b}_1 + \hat{b}_2 X_i$$

¹ La probabilité d'un événement se définit en effet comme la limite de sa fréquence relative lorsque la taille de l'échantillon tend vers l'infini.

Revenus	Nombre de familles	Nombre de familles possédant une maison
6	40	8
8	50	12
10	60	18
13	80	28
15	100	45
20	70	36
25	65	39
30	50	33
35	40	30
40	25	20

Exemple de données

Peut-on estimer ce modèle par les moindres carrés ? Pas sans précaution. On peut en effet montrer que les résidus de ce modèle (pour autant que le nombre d'observations par classe soit suffisant et que chaque observation dans chaque classe de revenu est le résultat d'un tirage aléatoire indépendant dans

une binomiale) suivent la distribution suivante : $u_i \cong N\left[0, \frac{1}{N_i P_i (1 - P_i)}\right]$. Ils sont donc

hétéroscédastiques. Il faut donc pour estimer le modèle passer par une procédure des moindres carrés pondérées (algorithme GLS). On utilisera comme coefficient de pondération la variance estimée des résidus. On procédera donc comme suit :

- pour chaque classe de la variable X, on calcule par probabilité que Y_i soit égal à 1 ($\hat{P}_i = \frac{n_i}{N_i}$ où n_i est le nombre d'observations pour lesquelles $Y_i = 1$ dans la classe i et N_i est le nombre total d'observation dans la classe i);
- pour X_i , on obtient $\hat{L}_i = \ln\left[\frac{\hat{P}_i}{(1 - \hat{P}_i)}\right]$;
- pour résoudre le problème d'hétéroscédasticité, on transforme le modèle de la manière suivante: $\sqrt{w_i} L_i = \mathbf{b}_1 \sqrt{w_i} + \mathbf{b}_2 \sqrt{w_i} X_i + \sqrt{w_i} u_i$ où les pondérations sont $w_i = N_i \hat{P}_i (1 - \hat{P}_i)$.
- on peut cette fois procéder à l'estimation du modèle par les moindres carrés. Attention toutefois, il s'agit bien d'un modèle sans constante.

Revenus	Nombre de familles	Nombre de familles possédant une maison	Pi	Li	Wi	Li*	Sart(wi)	Xi*
6	40	8	0.200	(1.386)	6.400	(3.507)	2.530	15.179
8	50	12	0.240	(1.153)	9.120	(3.481)	3.020	24.159
10	60	18	0.300	(0.847)	12.600	(3.008)	3.550	35.496
13	80	28	0.350	(0.619)	18.200	(2.641)	4.266	55.460
15	100	45	0.450	(0.201)	24.750	(0.998)	4.975	74.624
20	70	36	0.514	0.057	17.486	0.239	4.182	83.632
25	65	39	0.600	0.405	15.600	1.601	3.950	98.742
30	50	33	0.660	0.663	11.220	2.222	3.350	100.489
35	40	30	0.750	1.099	7.500	3.009	2.739	95.851
40	25	20	0.800	1.386	4.000	2.773	2.000	80.000

XLSTAT - Régression / Début le 09/11/98 à 1:50:45 PM

Données analysées :

Variable Y : classeur = Logit.xls / feuille = Sheet1 / plage = \$H\$4:\$H\$13

Variables quantitatives : classeur = Logit.xls / feuille = Sheet1 / plage = \$I\$4:\$J\$13

Variables qualitatives : Aucune

	Moyenne	Ecart type	Min	Max
Y	-0.379141606	2.672781883	-3.50707815	3.008673662
Variable X:	3.456000224	0.908988377	2	4.974937186
Variable X:	66.36330224	31.81113865	15.17893277	100.4888053

Coefficient de corrélation r : 0.9824

Coefficient de détermination r² : 0.9650

Coefficient de détermination ajusté : 0.9550

Paramètres de la régression et statistiques correspondantes :

	Valeur	Ecart-type	t de Student	Probabilité correspondante	Intervalle de confiance à 95%
Constante	0				
Variable X:	-1.593237788	0.1115	-14.2898	0.0001	-1.8503 -1.3361
Variable X:	0.078668569	0.0054	14.4412	0.0001	0.0661 0.0912