PRÉSENTATION GÉNÉRALE DU MODÈLE LOGIT

Maryse Marpsat et Alain Trognon

Une analyse des comportements court le risque de rester incomplète si on se limite à l'observation de tableaux croisés ventilant une pratique selon un ou plusieurs critères. En effet, divers effets de structure peuvent conduire à des interprétations erronées ; il est alors nécessaire d'isoler les effets propres de telle ou telle variable.

Pour ce faire, les tabulations croisées sont en général insuffisantes : même pour des enquêtes dont l'échantillon est grand, on se heurte très vite aux problèmes que pose le grand nombre de cases qui ne regroupent qu'un effectif très faible de ménages.

Pour aller plus loin, et tenter d'isoler l'effet spécifique d'un facteur "toutes choses égales par ailleurs", il faut faire des hypothèses et postuler des régularités statistiques.

*Quand le phénomène étudié est continu (exemple : le revenu ou son logarithme, la consommation ou son logarithme), la méthode appropriée est l'*analyse de la variance. *Cette méthode est une extension naturelle du modèle de régression par les moindres carrés ordinaires, ou MCO.*

Toutefois, dans une étude sur le comportement des ménages ou des individus, les pratiques étudiées sont le plus souvent de nature discrète, qualitative. Le recours à une analyse économétrique d'un type particulier est alors nécessaire pour isoler les effets propres (on parlera aussi de "séparation des effets", d'"effet d'une variable toutes choses égales par ailleurs", ou d'"effet d'une variable conditionnellement aux variables introduites dans le modèle").

Cette présentation est extraite d'un document de travail rédigé par M. Marpsat, A. Trognon, D. Verger et alii, n° F9110.

Les variables qualitatives

Différents types de variables qualitatives se rencontrent fréquemment dans nos enquêtes auprès des ménages :

Les variables dichotomiques

Ce sont des variables qui prennent deux valeurs, on dira aussi qui ont deux modalités, souvent notées 0 et 1.

Exemples :

- la possession d'un bien durable : 1 si le ménage possède le bien,
  0 s'il ne le possède pas ;

- la pratique d'une activité : 1 si l'individu pratique l'activité ;
  0 s'il ne la pratique pas.

Les variables polytomiques

Ce sont des variables qui prennent plus de deux valeurs ou modalités.

On distingue deux sortes de variables polytomiques, qui seront traitées différemment dans les modèles :

les variables polytomiques ordonnées

Les différentes modalités sont ordonnées dans un ordre "naturel" :

Exemple: faire du sport : 1. tous les jours ;
2. une ou plusieurs fois par semaine ;
3. plus rarement.

les variables non ordonnées

Exemple : parmi les distractions possibles le samedi soir, la personne interrogée préfère :

1. la télévision ;
2. le théâtre ;
3. le cinéma.

Dans cette note, on traitera essentiellement des variables à deux modalités, que l'on notera 0 et 1. Quelques indications seront données sur les autres cas, qui feront l'objet d'une note ultérieure.

Pourquoi des modèles particuliers ?

On ne peut pas utiliser la même méthode que dans le cas continu, puisqu'en particulier, la variable expliquée Y ne prenant que deux valeurs, la perturbation u suivrait obligatoirement une loi discrète, ce qui est incompatible avec les hypothèses habituelles de continuité et de normalité des résidus (voir Gouriéroux, 1989).

En effet, si on écrivait :

$$Y_i = X_i\beta + u_i \qquad \text{pour l'individu } i,$$

alors on aurait : $u_i = 1 - X_i\beta$ avec la probabilité p_i

et $u_i = - X_i\beta$ avec la probabilité $1 - p_i$

où $p_i = P [Y_i = 1]$, soit une loi discrète pour u_i.

Niveau d'utilité, variables latentes

Les méthodes utilisées partent du principe que le phénomène observé est la manifestation visible d'une *variable latente* Z inobservable qui, elle, est continue. On se ramène alors conceptuellement à un modèle d'*analyse de la variance* ou plus généralement à un modèle linéaire sur cette variable latente, le problème à résoudre étant celui de l'estimation de ce modèle.

Exemple de variable latente : dans le cas de la possession d'un bien durable, la variable latente peut être "l'intensité du désir" de posséder le bien : tant que cette intensité reste inférieure à un certain seuil, on observe $Y_i = 0$ (le ménage i ne possède pas le bien), quand elle le dépasse on observe $Y_i = 1$ (le ménage i possède le bien). On peut aussi

formuler le problème en terme de fonction d'utilité : pour le ménage i de caractéristiques X_i (âge, sexe de la personne de référence, revenu etc.), la possession du bien procure un *niveau d'utilité* $U(1, X_i)$, alors que la non possession procure un niveau $U(0, X_i)$. On a alors :

$$Y_i = 1 \Longleftrightarrow U(1, X_i) > U(0, X_i)$$

et

$$Y_i = 0 \Longleftrightarrow U(0, X_i) > U(1, X_i)$$

le ménage choisissant la situation qui lui procure le plus haut niveau d'utilité.

On se ramène au cas de la variable latente en posant :

$$Z_i = U(1, X_i) - U(0, X_i)$$

On a alors :

$$Y_i = 1 \Longleftrightarrow Z_i > 0$$
$$Y_i = 0 \Longleftrightarrow Z_i < 0$$

Il y a possession du bien lorsque la variable latente Z_i dépasse le seuil 0.

Le(s) modèle(s) théorique(s)

Notons Y la variable dichotomique à expliquer, dite aussi variable dépendante, dont on supposera qu'elle prend les valeurs 0 ou 1.

On observe les valeurs que prend Y sur un ensemble d'individus (ou de ménages) indiciés par i, $i = 1,..., I$. I est la taille de l'échantillon. Soit Z la *variable latente* sous-jacente au phénomène.

Le modèle postule une relation du type :

$$Z = X\beta + u$$

où X est un ensemble de variables dites exogènes ou explicatives, qui peuvent être :

- des variables continues : le revenu, l'âge (dont l'effet est alors linéaire, voir plus loin dans les spécifications du modèle) ;

- des variables "discrétisées" : le revenu en tranches, l'âge décennal (ce qui permet de mettre en évidence des effets non linéaires) ;

- des variables qualitatives : la CSP, la catégorie de commune.

Dans le cas de variables discrétisées ou qualitatives, il convient de choisir une situation de référence (voir plus loin).

La probabilité que l'individu i soit dans l'état $Y_i = 1$ est alors :

$$\begin{aligned} p_i = P[Y_i = 1] &= P[Z_i > 0] \\ &= P[X_i \beta > -u] \\ &= F(X_i \beta) \end{aligned}$$

si on note F la *fonction de répartition* de $-u$, c'est-à-dire la fonction définie par :

$$F(w) = P[-u < w].$$

Le choix du modèle porte sur le choix de F.

Deux fonctions sont couramment utilisées et seront traitées ici :

* F = fonction de répartition de la loi normale (modèle PROBIT) ;

* F = fonction de répartition de la loi logistique (modèle LOGIT).

Toutefois, d'autres fonctions peuvent être choisies. Ainsi, la procédure LOGISTIC de la version 6 de SAS, dont on traitera plus loin, permet également de prendre pour F la fonction de répartition de la loi de Gompertz.

Les modèles PROBIT et LOGIT

Le modèle PROBIT est celui pour lequel F est la fonction de répartition de la loi normale centrée réduite :

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$

ce qui donne :

$$P\ [Y = 1] = \int_{-\infty}^{X\beta} \frac{1}{\sqrt{2\pi}}\ e^{-t^2/2}\ dt = \Phi\,(X\beta)$$

Le modèle LOGIT est celui pour lequel F est la fonction de répartition de la loi logistique :

$$F\,(w) = L\,(w) = \frac{e^w}{1+e^w} = \frac{1}{1+e^{-w}}$$

ce qui donne:

$$P\,[Y = 1] = \frac{1}{1+\exp\,(-X\,\beta\,)} = L\,(X\beta\,)$$

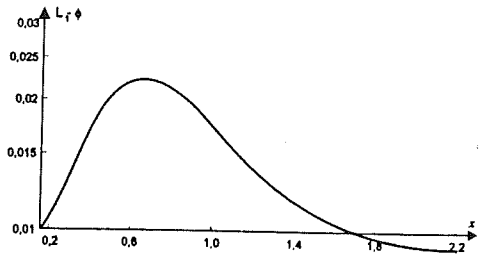
Comparaison des deux modèles

L (fonction de répartition de la loi logistique) et Φ (fonction de répartition de la loi normale) sont toutes les deux symétriques par rapport à 0, et comprises entre 0 et 1 (ce qui convient pour représenter une probabilité).

La loi logistique de fonction de répartition L a pour moyenne 0, pour variance $\frac{\pi^2}{3}$; il est donc naturel de comparer à $\Phi\,(w)$, fonction de répartition de N (0, 1), la fonction $L_1\,(w)$ où

$$L_1\,(w) = \frac{1}{1+\exp\,(-\pi\,w\,/\,\sqrt{3})}$$

La figure ci-dessous donne en fonction de w, la différence $L_1(w) - \Phi(w)$ des fonctions de répartition :



(référence : Gouriéroux 1989, p 29).

Ces lois étant proches, dans la plupart des cas pratiques on peut indifféremment choisir l'un ou l'autre modèle. Le modèle LOGIT a l'avantage d'une plus grande simplicité numérique, le modèle PROBIT est en revanche plus proche du modèle habituel de régression par les moindres carrés ordinaires, où la référence à des perturbations normales est souvent invoquée.

Attention toutefois lorsque vous voudrez comparer les estimateurs obtenus à partir des différents modèles. La PROC LOGISTIC utilise Φ et L (non pas L_1) : les estimateurs obtenus avec le modèle LOGIT seront donc $\pi/\sqrt{3}$ fois plus grands environ que ceux obtenus par le modèle PROBIT.

L'estimation : formules, précisions techniques

Le principe de la méthode

La méthode d'estimation adoptée est celle du maximum de vraisemblance.

L'enquête donne I observations indépendantes (Y_i, X_i).

Les Y_i sont des variables de Bernoulli $(1, p_i)$ où :

$$p_i = P[Y_i = 1]$$

La vraisemblance est alors :

- pour une observation : $p_i^{Y_i}(1 - p_i)^{1-Y_i} = L_i(\beta)$

– pour I observations : $\Lambda_I(\beta) = \prod_{i=1}^{I} L_i(\beta)$

soit :

$$\Lambda_I(\beta) = \prod_{i=1}^{I} F(X_i \beta)^{Y_i} [1 - F(X_i \beta)]^{1-Y_i}$$

La log-vraisemblance est :

$$L_I(\beta) = \log \Lambda_I(\beta) = \sum_{i=1}^{I} \text{Log } L_i(\beta)$$

soit :

$$L_I(\beta) = l = \sum_{i=1}^{I} Y_i \log F(X_i \beta) + \sum_{i=1}^{I} (1 - Y_i) \log [1 - F(X_i \beta)]$$

La procédure d'estimation consiste à rechercher la valeur $\hat{\beta}$ de β qui maximise la vraisemblance ou plus précisément son logarithme $L_I(\beta)$, qu'on notera dorénavant L.

Cas particulier du LOGIT

Dans le cas du LOGIT, on a :

$$\Lambda_I(\beta) = \prod_{i=1}^{I} \frac{1}{[1 + \exp(-X_i \beta)]^{Y_i}} \left[1 - \frac{1}{1 + \exp(-X_i \beta)}\right]^{1-Y_i}$$

$$= \prod_{i=1}^{I} \frac{[\exp(-X_i \beta)]^{1-Y_i}}{1 + \exp(-X_i \beta)}$$

$$L = \text{Log } \Lambda_I(\beta) = \sum_{i=1}^{I} (1 - Y_i)(-X_i \beta) - \sum_{i=1}^{I} \text{Log } [1 + \exp(-X_i \beta)]$$

De plus :

$$p_i = P[\,Y_i = 1\,] = \frac{1}{1 + \exp(-X_i \beta)}$$

Pour une combinaison particulière des variables explicatives, soit $X^* = (x_1^*, \ldots, x^*_p)$ on peut étudier la probabilité de choix prédite par le modèle :

$$\hat{p}^* = \frac{1}{1 + \exp(-X^* \hat{\beta})}$$

Il en est de même pour chaque individu *i*. La probabilité de choix (par exemple de choisir de posséder un bien) pour l'individu *i* est estimée, ou prédite, par :

$$\hat{p}_i = \frac{1}{1 + \exp(-X_i \hat{\beta})}$$

L'algorithme utilisé

Dans le cas des modèles LOGIT ou PROBIT, la log-vraisemblance L est concave et $\hat{\beta}$ est la solution de l'équation :

$$\frac{\partial L}{\partial \beta} = 0$$

c'est-à-dire :

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{I} \frac{Y_i - F(X_i \beta)}{F(X_i \beta)\,[\,1 - F(X_i \beta\,]}\, f(X_i \beta)\, X_i = 0$$

où f est la dérivée de F.

Cette solution est unique dans les cas usuels de non-dégénérescence.

Donc toute procédure itérative convergente (dont l'emploi pour résoudre l'équation différentielle est nécessaire car l'équation est non-linéaire) converge vers $\hat{\beta}$. La procédure employée dans la plupart des cas est basée sur l'algorithme de Newton-Raphson. La procédure employée par la PROC LOGISTIC de SAS est différente et utilise une méthode itérative de moindres carrés repondérés (Iteratively Reweighted

Least Squares, ou IRLS). Dans les deux cas, à partir d'une valeur initiale $\hat{\beta}^{(0)}$, on corrige l'estimation selon une formule du type :

$$\hat{\beta}^{(i+1)} = \hat{\beta}^{(i)} + c^{(i)}$$

jusqu'à obtenir la stabilité, en l'occurrence jusqu'au moment où la valeur absolue de la différence entre les valeurs calculées pour le logarithme de la vraisemblance à deux étapes successives soit en deçà d'un seuil fixé à l'avance. Pour la PROC LOGISTIC, toutefois, on considère que les itérations ont convergé lorsque la différence maximale entre les estimateurs des différents paramètres est inférieure à un seuil, par défaut 10^{-4}.

Pour plus de détails sur la méthode IRLS voir SAS/STAT User's guide, vol.2. Pour plus de détails sur la méthode de Newton-Raphson, voir Agresti 1990 ou Gouriéroux 1989.

Quelques propriétés asymptotiques[1] *de l'estimateur du maximum de vraisemblance*

Sous des hypothèses très générales, l'estimateur du maximum de vraisemblance a de bonnes propriétés. Il est asymptotiquement normal :

$$\hat{\beta} \xrightarrow[\text{asymp.}]{} N(\beta, V\hat{\beta})$$

où :

$$V\hat{\beta} = -\left(E \frac{\partial^2 L}{\partial \beta \partial \beta'}\right)^{-1}$$

Or :

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{I} \left[\frac{Y_i}{F^2(X_i\beta)} + \frac{1-Y_i}{[1-F(X_i\beta)]^2}\right] f^2(X_i\beta) X_i X'_i$$

(1) Asymptotique : lorsque I est grand

$$+ \sum_{i=1}^{I} \frac{Y_i - F(X_i \beta)}{F(X_i \beta)\,[1 - F(X_i \beta)]}\, f'(X_i \beta)\, X_i X'_i$$

dont l'espérance vaut :

$$E\,\frac{\partial^2 L}{\partial \beta\, \partial \beta'} = -\sum_{i=1}^{I} \frac{f^2(X_i \beta)}{F(X_i \beta)\,[1 - F(X_i \beta)]}\, X_i X'_i$$

La matrice de variance-covariance asymptotique de $\hat{\beta}$ vaut donc :

$$V\hat{\beta} = + \left[\sum_{i=1}^{I} \frac{f^2(X_i \beta)}{F(X_i \beta)\,[1 - F(X_i \beta)]}\, X_i X'_i \right]^{-1}$$

dont un estimateur est obtenu en calculant la valeur précédente au point $\hat{\beta}$.

Cas particulier du LOGIT :

Dans ce cas, on a :

$$F(w) = \frac{1}{1 + e^{-w}}$$

$$f(w) = F'(w) = \frac{e^{-w}}{(1 + e^{-w})^2} = \frac{1}{1 + e^{-w}} \cdot \frac{e^{-w}}{1 + e^{-w}}$$

$$f(w) = F(w)\,.\,[1 - F(w)]$$

et donc : $f(X_i \hat{\beta}) = F(X_i \hat{\beta})\,[1 - F(X_i \hat{\beta})] = \hat{p}_i (1 - \hat{p}_i)$

soit :

$$\hat{V}\hat{\beta} = \left[\sum_{i} X'_i X_i \hat{p}_i (1 - \hat{p}_i) \right]^{-1}$$

Cas particulier du PROBIT :

f est la densité de la loi normale centrée réduite et F son intégrale.

Les tests

(1) Test de la nullité d'un coefficient

On veut tester la nullité du coefficient β_j, c'est-à-dire de la $j^{ème}$ composante du vecteur de paramètres β.

β_j, est le coefficient correspondant à la $j^{ème}$ variable explicative ($j^{ème}$ colonne de la matrice X).

On considère la statistique de Student :

$\frac{\hat{\beta}_j}{\sqrt{\hat{v}\,\hat{\beta}_j}}$ où : $\hat{\beta}_j$ est la $j^{ème}$ composante de l'estimateur

$\hat{v}\,\hat{\beta}_j$ est le $j^{ème}$ coefficient de la diagonale de la matrice de variance-covariance estimée de $\hat{\beta}$

$\sqrt{\hat{v}\,\hat{\beta}_j}$ en est l'écart-type estimé (standard deviation)

On compare habituellement cette statistique au seuil de significativité à 5 % d'une loi normale (environ 2).

Dans la procédure LOGISTIC de SAS, la significativité de chaque coefficient β_j est testée à partir de la statistique de Wald :

$$W = \frac{\hat{\beta}_j^2}{\hat{v}\,\hat{\beta}_j}$$ soit le carré de la statistique de Student.

Cette statistique suit asymptotiquement une loi du chi2 à 1 degré de liberté. L'hypothèse de la nullité de β_j est rejetée lorsque la statistique de Wald dépasse un certain seuil, environ 4 pour une significativité à 5 %.

(2) Test d'une liaison de la forme $\sum_{k=1}^{K} \lambda_k \beta_k = C$

Si on note $\hat{V}\hat{\beta}$ la matrice de variance-covariance estimée de l'estimateur $\hat{\beta}$ et Q' le vecteur ligne ($\lambda_1 \ldots \lambda_K$), on a le résultat asymptotique suivant :

$$\frac{Q'\hat{\beta} - c}{\sqrt{Q'(\hat{V}\hat{\beta})Q}} \xrightarrow[\text{asymp.}]{} N(0, 1) \text{ si l'hypothèse nulle } Q'\beta = C \text{ est vraie}$$

Si l'hypothèse alternative du test est $Q'\beta \neq C$, l'hypothèse "nulle" est rejetée si la valeur absolue de la statistique précédente dépasse un certain seuil de significativité.

Le cas 1 est bien sûr un cas particulier de 2, lorsque seul λ_j est non nul et C = 0.

Test de la nullité d'un ensemble de coefficients

On peut souhaiter tester la nullité d'un ensemble de q coefficients (par exemple tous ceux concernant les différentes variables introduites pour représenter une dimension explicative (cf infra) telle que la CSP, ou bien le revenu en tranches, ou bien l'âge quinquennal etc.). On peut souhaiter tester également la nullité de l'ensemble des coefficients.

L'hypothèse de la nullité d'un ensemble de q coefficients s'écrit sous la forme $Q'\beta = 0$, où Q' est une matrice diagonale où seuls les coefficients correspondant aux β_j dont on veut tester la nullité sont égaux à 1, les autres étant nuls. Par exemple, dans le cas où $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$ et où on veut tester

$\beta_1 = 0$ et $\beta_2 = 0$, on aura : $Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

On dispose de plusieurs tests, par exemple :

- le test de Wald

– le test du rapport de vraisemblance

** Test de Wald*

$$W = (Q'\hat{\beta})' \, [Q' \, (\hat{V}\hat{\beta}) Q]^{-1} \, (Q'\hat{\beta}) \xrightarrow[\text{asymp.}]{} \chi^2_q$$

W tend asymptotiquement vers un chi 2 à q degrés de liberté où q est le rang de la matrice Q. Rappelons que dans le cas d'une variable catégorielle à p modalités, comme les CSP, l'une des modalités est prise comme niveau de référence et son coefficient est donc nul. La statistique de Wald sur les coefficients des modalités qui restent sera donc convergente asymptotiquement vers un chi 2 à p-1 degrés de liberté.

Ici encore, l'hypothèse "nulle" $Q' \beta = 0$ sera rejetée lorsque la valeur de la statistique de Wald dépassera un seuil critique.

** Test du rapport de vraisemblance*

Si L désigne la log-vraisemblance, $\hat{\beta}$ l'estimateur du maximum de vraisemblance, $\hat{\beta}_0$ l'estimateur du maximum de vraisemblance sous la contrainte $Q' \beta = 0$, on a :

$$LRT = 2 \, [L(\hat{\beta}) - L(\hat{\beta}_0)] \xrightarrow[\text{asymp.}]{} \chi^2_q$$

Ici aussi, l'hypothèse de nullité simultanée des coefficients considérés doit être rejetée si la valeur de la statistique dépasse un seuil critique.

Application : choix entre 2 modèles dont l'un est une version "réduite" de l'autre.

Modèle 1 : les variables explicatives sont $X_1, \ldots, X_p$

Modèle 2 : $X_1, \ldots, X_p, X_{p+1}, \ldots, X_{p+k}$

Préférer 1 à 2, c'est accepter l'hypothèse que, dans le second modèle, les *k* coefficients $\beta_{p+1}, \ldots, \beta_{p+k}$ sont nuls. Cette hypothèse s'écrit sous la forme $Q' \beta = 0$ comme on l'a déjà vu.

On choisira le modèle 2 si :

$$LRT = 2 \, [L(\hat{\beta}) - L(\hat{\beta}_0)]$$

est supérieur à la valeur critique au seuil de *a* % du chi 2 à *k* degrés de liberté.

* *Attention* : ce type de choix entre 2 modèles dont l'un est une version réduite de l'autre se présente en particulier dans le cas d'estimations Backward (on retire des variables au modèle selon certains critères de choix), Forward (on en ajoute), ou Stepwise (alternativement, on retire et on ajoute des variables au modèle). Toutefois, la PROC LOGISTIC de SAS 6 choisit entre les modèles en utilisant la statistique du score.

Cette statistique est une forme quadratique construite à partir du vecteur des dérivées partielles de la log-vraisemblance par rapport au vecteur de paramètres β et évaluée en β_0 (c'est-à-dire sous l'hypothèse nulle).

On a alors :

$$S = \left(\frac{\partial L}{\partial \beta} \right)' (\beta_0)\ I^{-1}(\beta_0) \frac{\partial L}{\partial \beta}(\beta_0)$$

(où I est l'information de Fisher), qui suit asymptotiquement une loi du chi 2 à *k* degrés de liberté.

On choisira alors le modèle 2 (c'est-à-dire celui qui comporte le plus de variables explicatives) lorsque S sera supérieur à la valeur critique au seuil de *a* % du ch 2 à *k* degrés de liberté. SAS édite la "*p*-value" de la statistique S dite aussi "statistique du chi 2 résiduel", c'est-à-dire la probabilité que sous l'hypothèse nulle (modèle 1) la statistique S dépasse la valeur observée. Cette "*p*-value" doit être faible pour choisir le modèle 2.

Cas plus général : test d'une hypothèse linéaire de la forme $Q'\beta = C$

où . Q' est une matrice de coefficients constants connus de dimension $q \times K$ (K nombre de variables dans le modèle estimé, y compris la constante)

. C est un vecteur de constantes connues, déterminées par l'utilisateur.

Les *q* lignes de Q sont linéairement indépendantes. On voit que les cas traités précédemment sont tous des cas particuliers de celui-ci.

On peut ici encore utiliser le test de Wald :

$$W = (Q'\hat{\beta} - C)' \, [Q' \, (\hat{V}\hat{\beta}) \, Q]^{-1} \, (Q'\hat{\beta} - C) \xrightarrow[\text{asymp.}]{} \chi^2_q$$

ou celui du rapport de vraisemblance :

$$LRT = 2 \, [L(\hat{\beta}) - L(\hat{\beta}_0)] \xrightarrow[\text{asymp.}]{} \chi^2_q$$

où $\hat{\beta}_0$ est l'estimateur obtenu en maximisant la vraisemblance sous la contrainte $Q'\beta = C$

Comme précédemment, l'hypothèse $Q'\beta = C$ doit être rejetée si la valeur de la statistique dépasse un certain seuil.

Tests de la validité générale du modèle

Existe-t-il des statistiques permettant de juger de la bonne adéquation du modèle, en jouant un rôle analogue à celui du R^2 classique ? Les auteurs en ont proposé plusieurs, souvent critiquables à un titre ou à un autre[1].

Voici celles fournies par la PROC LOGISTIC :

- **le rapport de vraisemblance** (l'hypothèse nulle étant celle où le modèle contient la seule constante) ;

- **la statistique du score** (ou du chi 2 résiduel) déjà définie plus haut ;

- **le critère d'Akaike**
  $AIC = -2 \log L + 2K$
  où K est le nombre de paramètres à estimer ;

- **le critère de Schwartz**
  $SC = -2 \log L + K \, \text{Log} \, I$
  où I est le nombre total d'observations.

(1) Il est en particulier difficile d'apporter les corrections adéquates pour comparer des modèles ayant des nombres de degrés de liberté différents.

Les critères de Schwartz et d'Akaike sont utiles pour comparer des modèles différents portant sur les mêmes données. On préférera le modèle pour lequel ces statistiques ont la valeur la plus faible.

D'autres approches permettent d'évaluer la capacité prédictive du modèle :

- **les tables de classification** (voir l'option CTABLE dans l'instruction MODEL). On "prédit" Y_i par $\hat{Y}_i$ de la façon suivante :

$$\hat{Y}_i = 1 \text{ si la probabilité estimée de valoir 1 dépasse 1/2, } \hat{Y}_i = 0 \text{ sinon.}$$

La "sensibilité" (sensitivity) est la proportion de vraies valeurs 1 qui sont prédites valoir 1. La "spécificité" (specificity) répond à la définition analogue pour les valeurs 0. Le "taux d'erreur par excès" (false positive rate) est la proportion de prédictions 1 qui valent en réalité 0. Le "taux d'erreur par défaut" (false negative rate) la proportion de prédictions 0 qui valent en réalité 1.

Le seuil de 1/2 utilisé pour prédire $\hat{Y}_i = 1$ peut être modifié. En effet dans le cas de pratiques très ou très peu répandues, et ne présentant que de faibles disparités, la prédiction est soit toujours 0, soit toujours 1. Ainsi, considérons la possession de fer à repasser ; 94 % des ménages en possèdent un ; toutefois, un critère comme l'âge a une influence certaine quoique limitée. Or ces tables de classification ne permettent pas de choisir entre un modèle explicatif comprenant la constante seule et le modèle qui introduit l'âge : on prédira toujours qu'il y a possession.

- **les prédictions et observations concordantes**.

On considère toutes les paires d'observations ayant des valeurs observées de Y différentes, soient 1 et 0. Parmi ces paires, on compte celles pour lesquelles l'observation où Y = 1 a une probabilité estimée que Y = 1 plus grande que l'observation où Y = 0. On dit alors que la paire est concordante. Elle est discordante lorsque la probabilité que Y = 1 est plus faible pour l'observation où Y = 1 que pour celle où Y = 0.

Les paires qui ne sont ni concordantes ni discordantes sont dites "liées" (tied) ou "ex-æquo".

Si I est le nombre total d'observations, t le nombre de paires ayant des valeurs observées de Y différentes, n_c le nombre de paires concordantes, n_d le nombre de paires discordantes, $t - n_c - n_d$ le nombre de paires "liées", SAS calcule quatre indices de "corrélation du rang" (rank correlation) :

$$c = [\, n_c + 0{,}5\, (\, t - n_c - n_d\,)\,] \,/\, t$$

Somer's D = ($n_c - n_d$) / t

Goodman -Kruskal Gamma = ($n_c - n_d$) / ($n_c + n_d$)

Kendall's Tau – a = ($n_c - n_d$) / (0,5 I [I – 1])

Ces quatre indices sont en quelque sorte des mesures d'association entre la probabilité prédite et la valeur de la variable explicative. Cette association est d'autant plus forte (et on est d'autant plus satisfait) que les indices sont élevés, c'est-à-dire proches de 1. En effet tous ces indices sont croissants lorsque n_c croît, décroissants lorsque n_d croît et varient entre les bornes suivantes :

C : entre 0 et 1

Somer's D : entre -1 et +1

Gamma : entre -1 et +1

Kendall's Tau-a : entre -1 et +1

Le cas extrême où l'indice prend la valeur +1 est celui où la totalité des paires ayant pour un élément Y = 0 et pour l'autre Y = 1 sont concordantes (c'est-à-dire que la probabilité estimée que Y = 1 est plus forte pour l'observation telle que Y = 1) : la prévision correspond "au mieux" à la réalité.

Mise en œuvre de la procédure LOGISTIC de la version 6 de SAS

Quelques remarques et mises en garde préalables

- La procédure LOGISTIC ajuste des modèles à résidus logistiques (LOGIT) ou normaux (PROBIT) ou encore correspondant à la loi de Gompertz (voir plus haut). Dans la procédure, la variable dépendante doit être soit dichotomique (ce qui est le cas traité dans cette note), soit polytomique ordonnée. Dans ce dernier cas les hypothèses faites sur le modèle sont assez restrictives. La procédure CATMOD traite aussi les modèles dichotomiques et polytomiques ordonnés. Elle traite en plus les modèles polytomiques non ordonnés. Mais elle est plus complexe et la procédure LOGISTIC est conseillée pour les modèles dichotomiques et polytomiques ordonnés.

- La procédure LOGISTIC traite les variables explicatives comme si elles étaient continues. Il convient donc de dichotomiser les variables explicatives qualitatives, telles que CSP, sexe, mais aussi tranches de revenu ou d'âge.

Ici quelques remarques générales concernant la dichotomisation des variables explicatives qualitatives.

1. La variable explicative X est déjà à deux modalités, 0 et 1 : on ne change rien. On fera figurer X dans la liste des variables explicatives et la procédure considérera que 0 est la modalité de référence.

2. La variable explicative X est à deux modalités quelconques (par exemple : 8 et 9). Si on choisit 8 pour modalité de référence, on fera figurer dans la liste des variables explicatives X1 définie au préalable par :

$$X1 = (X = 9) ;$$

3. La variable explicative X a n modalités prenant les valeurs 1,..., n.

On utilisera l'instruction ARRAY.

Exemple : la catégorie socioprofessionnelle est la variable PPCS qui vaut de 1 à 8.

On écrira :

```
ARRAY P(J)  PPCS1 - PPCS8 ;

DO  J = 1  TO  8 ;

P = (PPCS = J) ;

END ;
```

4. La variable explicative X a $n + 1$ modalités prenant les valeurs 0, 1,..., n.

Première solution : on recodifie au préalable par X=X + 1 et on se ramène au cas précédent.

Deuxième solution : prenons l'exemple du diplôme de la personne de référence, DIPLOPR, qui varie de 0 à 5. On écrira :

```
ARRAY  D(M)   DIPR0-DIPR5 ;

DO  M = 1   TO  6 ;
```

D = (DIPLOPR = M-1) ;

5. On veut à la fois dichotomiser et regrouper des modalités.

Exemple : le revenu du ménage est indiqué par la variable REVENU qui prend des valeurs de 1 à 8 (8 tranches). On veut opérer les regroupements suivants :

1 et 2, 3 et 4, 5 à 7, 8.

On écrira :

REV1 = (REVENU = 1 ! REVENU = 2) ;

REV2 = (REVENU = 3 ! REVENU = 4) ;

REV3 = (5 < = REVENU < = 7) ;

REV4 = (REVENU = 8) ;

Ne pas oublier qu'il faut une modalité de référence (sinon PROC LOGISTIC prend la dernière). Cette modalité est celle qui est omise dans la liste des variables de l'instruction MODEL.

- La procédure LOGISTIC ajuste le modèle sur la probabilité de la modalité la plus faible.

Si donc vous avez codé :

$$Y = \begin{cases} 0 & \text{je ne possède pas un bien} \\ 1 & \text{je le possède} \end{cases}$$

les coefficients de la régression seront positifs pour les modalités explicatives correspondant à une moindre possession du bien. Vous avez la possibilité, soit de changer le signe de vos coefficients lorsque vous donnez vos résultats, soit de recodifier au début du programme :

$$Y_1 = \begin{cases} 1 & \text{je possède le bien} \\ 2 & \text{je ne le possède pas} \end{cases}$$

Si vous ne recodifiez pas, cela revient à avoir tous les coefficients multipliés par (-1).

En effet :

$$\hat{p}(Y=0) = \frac{1}{1+\exp(-X\hat{\beta}_{(0)})} = \frac{\exp(X\hat{\beta}_{(0)})}{\exp(X\hat{\beta}_{(0)})+1} = 1 - \frac{1}{1+\exp(X\hat{\beta}_{(0)})}$$

D'autre part :

$$\hat{p}(Y=0) = 1 - \hat{p}(Y=1) = 1 - \frac{1}{1+\exp(-X\hat{\beta}_{(1)})}$$

Donc : $\hat{\beta}_{(0)} = -\hat{\beta}_{(1)}$

- Dans le cas de données de départ très nombreuses on peut effectuer un sondage ou bien utiliser la syntaxe dite "événements/expériences" (events/trials) de la PROC LOGISTIC. Cette syntaxe ne sera pas exposée ici.

Quelques rappels de syntaxe

(pour plus de détails, voir la brochure SAS intitulée SAS/STAT User's Guide volume 2 version 6, ou la traduction - provisoire - de Maryse Marpsat, disponible au secrétariat de la division Études Sociales).

Instructions obligatoires
- PROC LOGISTIC < options 1 > ;
- MODEL Y = X1 X2 ... </options 2 > ;

Instructions facultatives
- BY variables ;
- OUTPUT < OUT = table sas >
- < mot–clé = nom1 mot–clé = nom2 ... >
- </ALPHA = valeur > ;
- WEIGHT variable ;

Les parties entre < > sont optionnelles.

Parmi les options 1 :

. DATA= pour préciser la table SAS où sont les données de départ ;

. des options pour modifier les impressions automatiques.

. OUTEST= crée une table SAS qui contient les estimateurs définitifs des paramètres et en option leur covariance estimée. Dans le cas d'un modèle dichotomique, les noms des variables dans cette table sont les mêmes que ceux des variables explicatives de MODEL plus le nom INTERCEP pour l'estimateur de la constante.

Parmi les options 2 :

. LINK= permet de traiter le modèle PROBIT (LINK= NORMIT) et celui lié à la loi de Gompertz (LINK= CLOGLOG). Par défaut, LINK= LOGIT.

. NOINT ajuste un modèle sans terme constant

. SELECTION= pour sélectionner la méthode de construction du modèle. Par défaut SELECTION= NONE (l'ajustement se fait sur toutes les variables explicatives indiquées). On peut adopter SELECTION= BACKWARD, SELECTION= FORWARD, SELECTION= STEPWISE.

D'autres options précisent les impressions désirées quand SELECTION= est précisé, l'état de départ, les niveaux de significativité désirés pour qu'une variable soit retenue...

. CTABLE imprime une table de classification (voir plus haut), pour laquelle le seuil retenu par défaut est 0,5. Ce seuil peut être modifié par l'option PPROB=

. Divers diagnostics sur la régression. En particulier, IPLOTS donne des graphiques représentant pour chaque observation la valeur d'un certain nombre de statistiques. Attention quand vous avez beaucoup d'observations !

. MAXITER= permet de modifier le nombre d'itérations (cf infra). Le nombre par défaut est 25.

Un exemple de sortie interprétée

La variable dépendante est ALVAC, où :

ALVAC = 1 si l'individu prend des vacances

ALVAC = 2 s'il n'en prend pas

Les variables explicatives sont :

TTU0, TTU1, TTU2 : catégorie de commune

PPCS1 à PPCS8 : CSP de la personne de référence du ménage

TA12, TAS5 : taille du ménage (1 ou 2 personnes, 5 et plus)

REV1, REV2 : tranche de revenu du ménage

DIPR0, DIPR1, DIPR23 : diplôme de la personne de référence du ménage

AGK1, AGK2, AGK3 : âge de l'individu (kish)

Auxquelles s'ajoute la constante (intercept) représentée par INTERCPT.

Nous avons éliminé de la sortie quelques statistiques descriptives élémentaires portant sur les variables explicatives, peu intéressantes quand il s'agit de variables dichotomiques. Pour cela nous avons utilisé l'option NOSIMPLE.

La moyenne - mean - combinée avec le nombre d'observations du fichier dans un modèle non pondéré permet toutefois de retrouver les effectifs de chaque modalité des variables explicatives. Il paraît plus simple de faire toujours précéder le modèle d'une PROC FREQ sur les modalités des variables explicatives. Bien qu'il n'y ait pas de limite inférieure à respecter sur ces effectifs, il conviendra d'être prudent quant à l'interprétation des coefficients estimés sur des strates d'effectif réduit (de l'ordre de moins de 20).

Le programme était donc :

```
PROC LOGISTIC NOSIMPLE ;

MODEL ALVAC = TTU0 TTU1 TTU2

              PPCS1 PPCS2 PPCS3 PPCS4

              PPCS6 PPCS7 PPCS8

              TA12 TAS5

              REV1 REV2

              DIPR0 DIPR1 DIPR23

              AGK1 AGK2 AGK3 / CTABLE ;
```

The LOGISTIC Procedure

Exemple 1a. Logit

Variable dépendante → Data Set: WORK.ESSAI
Response Variable: ALVAC
Response Levels: 2
Number of Observations: 10472
modèle logit → Link Function: Logit

Response Profile

Ordered Value	ALVAC	Count
1	1	8348
2	2	2124

Criteria for Assessing Model Fit ← Critères permettant de juger de l'ajustement du modèle

Criterion	① Intercept Only	② Intercept and Covariates	Chi-Square for Covariates
Critère d'Akaike → AIC	10563.976	8067.051	.
Critère de Schwartz → SC	10571.232	8219.437	.
−2 log-vraisemblance → −2 LOG L	10561.976	8025.051	2536.925 with 20 DF (p=0.0001) ③
Score	.	.	2415.300 with 20 DF (p=0.0001)

① modèle avec constante seule
② modèle avec constante et variables explicatives

③ Valeur de : $-2\log L_{①} - [-2\log L_{②}]$
La probabilité que le chi-2 à 20 degrés de liberté dépasse cette valeur est de p = 0.0001
L'hypothèse nulle [modèle sans variables explicatives] est donc rejetée

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	3.3039	0.2452	181.5900	0.0001	.
TTU0	-1.4058	0.1091	165.9573	0.0001	-0.345011
TTU1	-0.9756	0.1120	75.8509	0.0001	-0.226877
TTU2	-0.4178	0.1122	13.8707	0.0002	-0.107473
PPCS1	-1.2305	0.1513	66.1069	0.0001	-0.130454
PPCS2	-0.1167	0.1654	0.4977	0.4805	-0.015011
PPCS3	1.2003	0.2931	16.7676	0.0001	0.193451
PPCS4	0.4538	0.1658	7.4910	0.0062	0.009839
PPCS6	-0.4660	0.1199	15.1040	0.0001	-0.111739
PPCS7	-0.6459	0.1341	23.2093	0.0001	-0.151091
PPCS8	-0.8823	0.1521	33.6292	0.0001	-0.106345
TA12	0.1511	0.0742	3.9297	0.0474	0.041400
TA55	-0.3043	0.0900	11.4380	0.0007	-0.050156
REV1	-1.0367	0.0799	168.5154	0.0001	-0.206601
REV2	-0.6345	0.0682	86.5667	0.0001	-0.153394
DIPR0	-1.4825	0.1820	66.3574	0.0001	-0.356918
DIPR1	-0.9592	0.1041	27.1370	0.0001	-0.214854
DIPR23	-0.5164	0.1812	8.1213	0.0044	-0.140176
AGK1	1.3529	0.1433	89.1926	0.0001	0.219010
AGK2	1.0472	0.1190	77.4804	0.0001	0.224799
AGK3	0.6945	0.0926	56.2740	0.0001	0.191410

④ $\pi_j = \dfrac{\sigma_{\text{part. rép. loi logistique}}}{\sigma_{j\text{ème var. explic. sur l'échantillon}}}$

("loi logistique" serait remplacé par "loi normale" si LINK = NORMIT)

L'estimateur standardisé permet de comparer les modèles LOGIT et PROBIT.

↑ estimateur du paramètre $\hat{\beta}_j$

↑ écart-type de $\hat{\beta}_j$ $= \hat{\sigma}_j$

↑ stat. de Wald $\dfrac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} > l^2 = 4$ $\Rightarrow \beta_j$ non nul

↑ $p < 0.05$ $\Rightarrow \beta_j$ significativement non nul

④ ↑ estimateur standardisé $\dfrac{\hat{\beta}_j}{\pi_j}$

The LOGISTIC Procedure

proportion des paires concordantes et discordantes (voir texte sur les tests et autres indicateurs de validité du modèle)

Association of Predicted Probabilities and Observed Responses

Concordant	= 82.6%	Somers' D	= 0.656
Discordant	= 17.0%	Gamma	= 0.659
Tied	= 0.4%	Tau-a	= 0.212
(17731152 pairs)		c	= 0.820

Classification Table ⑤

table de classification (voir texte sur les tests)

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	7876	472	8348
	NO EVENT	1356	768	2124
	Total	9232	1240	10472

Sensitivity= 94.3% Specificity= 36.2% Correct= 82.5%
False Positive Rate= 14.7% False Negative Rate= 38.1%

NOTE: An EVENT is an outcome whose ordered response value is 1.

⑤ $a = 7876$ observations ont une valeur 1 pour ALVAC, réelle et prédite
$b = 472$ ont la valeur réelle 1, la valeur prédite 2
$c = 1356$ ont la valeur réelle 2, la valeur prédite 1
$d = 768$ ont la valeur réelle 2, la valeur prédite 2

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d

$\text{Sensitivity} = \frac{a}{a+b}$ $\text{Specificity} = \frac{d}{c+d}$ $\text{Correct} = \frac{a+d}{a+b+c+d}$

$\text{False positive rate} = \frac{c}{a+c}$

$\text{False negative rate} = \frac{b}{b+d}$

"Correct" désigne le % d'observations où la valeur réelle est égale à la valeur prédite.

Le fichier en sortie

Pour obtenir un fichier (une table SAS) en sortie, il faut faire appel à une instruction facultative, l'instruction OUTPUT.

On écrira alors :

```
PROC LOGISTIC ;

MODEL variable dépendante = variables explicatives ;

OUTPUT  OUT = nom de la table SAS en sortie

        <mot-clé  =  nom ... mot-clé  =  nom> ;
```

Le fichier en sortie est une nouvelle table SAS qui contient toutes les variables de la table en entrée. En option, l'instruction OUTPUT crée l'estimateur $X\hat{\beta}$ de la partie linéaire du modèle, son écart-type estimé, la probabilité estimée pour chaque individu d'avoir la modalité la plus faible $Y = i$, l'intervalle de confiance pour cette probabilité, et des statistiques d'aide au diagnostic sur la régression.

Pour obtenir en sortie, par exemple, la probabilité estimée (que l'individu ait pour Y la valeur la plus faible) on emploiera le mot-clé PREDICTED, ou P. Si on veut lui donner le nom EQUIP, on écrira :

```
PROC LOGISTIC ;

MODEL .... ;

OUTPUT  OUT = ...  P = EQUIP ;
```

TOUTEFOIS pour obtenir la probabilité p estimée il est plus simple d'utiliser l'option OUTEST de l'instruction PROC LOGISTIC et de calculer p pour les modalités qui nous intéressent.

Exemple : si on a : ALVAC = 1 partir en vacances

ALVAC = 2 ne pas partir

```
PROC LOGISTIC  OUTEST = TAB ;

MODEL ALVAC = TTU0   TTU1   TTU2
```

```
PPCS1    PPCS2  PPCS3  PPCS4
         PPCS6  PPCS7  PPCS8
         TA12   TAS5
         REV1   REV2
         DIPR0  DIPR1  DIPR23
         AGK1   AGK2   AGK3 / CTABLE ;
OUTPUT OUT = LOISIR1  P = PHAT ;
```

On veut obtenir la probabilité estimée de partir en vacances pour les individus pour lesquels les variables TTU0, PPCS1, TA5, REV2, DIPR0, AGK1 valent 1.

Deux possibilités :

```
1. DATA  A ; SET  TAB ;
X1 = - (INTERCEP + TTU0 + PPCS1 + TA5 + REV2 + DIPR0 + AGK1) ;
PHAT1 = 1/(1 + EXP[X1]) ;
PROC  PRINT  DATA = A ; VAR  PHAT1 ;
2. DATA B ; SET  LOISIR1 ;
IF  TTU0 = 1 & PPCS1 = 1 & TA5 = 1 & REV2 = 1 & DIPR0 = 1 & AGK1 = 1 ;
PROC  PRINT  DATA = B  (OBS = 1) ; VAR  PHAT ;
```

Les valeurs qu'on obtient par PHAT et PHAT1 sont égales et représentent p.

Pour la question de l'utilisation des pondérations lors du calcul de probabilités estimées, voir "Pondérer ou ne pas pondérer", dans la partie IX^{e} du document de travail F9110.

The LOGISTIC Procedure

Exemple 1b. Probit.

Data Set: WORK.ESSAI
Response Variable: ALVAC
Response Levels: 2
Number of Observations: 10472
Modèle PROBIT ← Link Function: Normit

Response Profile

Ordered Value	ALVAC	Count
1	1	8348
2	2	2124

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	10563.976	8068.454	.
SC	10571.232	8220.840	.
-2 LOG L	10561.976	8026.454	2535.522 with 20 DF (p=0.0001)
Score	.	.	2415.388 with 20 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1.0215	0.1255	210.6802	0.0001	.
TTU0	-0.7916	0.0582	184.7172	0.0001	-0.353194
TTU1	-0.5398	0.0597	81.6779	0.0001	-0.227667
TTU2	-0.2202	0.0592	13.8208	0.0002	-0.102749
PPCS1	-0.7208	0.0857	70.7500	0.0001	-0.147115
PPCS2	-0.0558	0.0890	0.3926	0.5309	-0.013017
PPCS3	0.4917	0.1280	14.7616	0.0001	0.143738
PPCS4	0.1059	0.0821	5.1313	0.0235	0.066738
PPCS6	-0.2640	0.0637	17.1605	0.0001	-0.114823
PPCS7	-0.3593	0.0733	24.0564	0.0001	-0.153253
PPCS8	-0.4921	0.0850	33.5410	0.0001	-0.107578
TA12	0.0849	0.0424	4.0059	0.0453	0.042195
TAS5	-0.1694	0.0504	11.2796	0.0008	-0.058707
REV1	-0.6014	0.0461	170.0346	0.0001	-0.217374
REV2	-0.3583	0.0386	85.9851	0.0001	-0.157104
DIPR0	-0.7992	0.0887	81.1652	0.0001	-0.348991
DIPR1	-0.4936	0.0900	30.0974	0.0001	-0.200533
DIPR23	-0.2478	0.0870	8.1125	0.0044	-0.122017
AGK1	0.7777	0.0795	95.6679	0.0001	0.229185
AGK2	0.6062	0.0672	81.3001	0.0001	0.236027
AGK3	0.4311	0.0540	63.7467	0.0001	0.215506

The SAS System 12:24 Thursday, July 25, 1991 2

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 82.6%		Somers' D =	0.656
Discordant = 17.0%		Gamma =	0.658
Tied = 0.4%		Tau-a =	0.212
(17731152 pairs)		c =	0.828

Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	7809	459	8348
	NO EVENT	1371	753	2124
	Total	9260	1212	10472

Sensitivity= 94.5% Specificity= 35.5% Correct= 82.5%
False Positive Rate= 14.8% False Negative Rate= 37.9%

NOTE: An EVENT is an outcome whose ordered response value is 1.

The SAS System 12:24 Thursday, July 25, 1991 2

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 82.6%		Somers' D =	0.656
Discordant = 17.0%		Gamma =	0.658
Tied = 0.4%		Tau-a =	0.212
(17731152 pairs)		c =	0.828

Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	7809	459	8348
	NO EVENT	1371	753	2124
	Total	9260	1212	10472

Sensitivity= 94.5% Specificity= 35.5% Correct= 82.5%
False Positive Rate= 14.8% False Negative Rate= 37.9%

NOTE: An EVENT is an outcome whose ordered response value is 1.

Modèle LOGIT, modèle PROBIT

Le modèle PROBIT est traité dans SAS, on l'a vu, en ajoutant l'option LINK= NORMIT à l'instruction MODEL.

Les résultats obtenus ne sont pas directement comparables. Il faut comparer les estimateurs standardisés (standardized estimates) qui tiennent compte de la différence de variance entre les deux distributions.

Le programme était ici :

```
PROC LOGISTIC NOSIMPLE ;

MODEL ALVAC= les mêmes variables

                    / CTABLE  LINK= NORMIT ;
```

Le lecteur se convaincra aisément que même si certains coefficients standardisés diffèrent légèrement, les conclusions qu'ils permettent de tirer sont identiques.

BIBLIOGRAPHIE

LOGIT (modèle dichotomique ou polytomique ordonné)

Quelques articles assez anciens mais donnant des éléments d'explication sur la modélisation :

Daniel VERGER, "L'achat d'un logement ne va pas sans achats d'équipements", *Économie et Statistique*, n° 161, décembre 1983.

Alain TROGNON, "Modèle de diffusion d'une innovation : l'exemple de la télévision couleur", *Annales de l'Insee*, n° 29, janvier 1978.

Daniel DEPARDIEU, Stéfan LOLLIVIER, "Les facteurs de l'absentéisme", *Économie et Statistique*, n° 176, avril 1985.

Stéfan LOLLIVIER, Daniel VERGER, "Les comportements en matière d'épargne et de patrimoine", *Économie et Statistique*, n° 202, septembre 1987. (logit polytomique univarié ordonné).

Quelques articles récents :

Françoise DUMONTIER, François de SINGLY, Claude THELOT,"La lecture moins attractive qu'il y a vingt ans", *Économie et Statistique*, n° 233, juin 1990, pp. 77 à 80 surtout.

Bénédicte REYNAUD, "Les modes de rémunération et le rapport salarial", *Économie et Prévision*, n° 92-93, 1990.

Claire SARMA, "Un exemple d'application du modèle logit : l'investissement immobilier des couples non mariés", *Économie et Prévision*, n° 91, 1989.

Michel BUA, Philippe GIRARD, Thierry PUJOL, Philippe REDONDO, "La politique de distribution des dividendes (1982-1986)", *Économie et Prévision*, n° 88-89, 1989.

Pascal BOUYAUX, "Une difficulté d'interprétation de l'approche LOGIT : l'exemple de l'économie des transports", *Économie et Prévision*, n° 91, 1989.

Michel GOLLAC, "L'ordinateur dans l'entreprise reste un outil de luxe", *Économie et Statistique*, n° 224, septembre 1989.

Pascal GARRIGUES, "Une France un peu plus sportive qu'il y a vingt ans... grâce aux femmes", *Économie et Statistique*, n° 224, septembre 1989.

Caroline ROY, "La gestion du temps des hommes et des femmes, des actifs et des inactifs", *Économie et Statistique*, n° 223, juillet-août 1989.

Jean-Yves FOURNIER, "Les absences au travail : 16 jours par an pour un ouvrier, 3,5 jours pour un cadre", *Économie et Statistique*, n° 221, mai 1989.

Pierre LAULHE, "1980-1985 : Les difficultés de l'insertion", *Économie et Statistique*, n° 216, décembre 1988.

Claire SARMA, "Les couples non mariés possèdent moins de patrimoine que les couples mariés" *Économie et Statistique*, n° 214, octobre 1988.

Stéfan LOLLIVIER, "Activité et arrêt d'activité féminine", *Économie et Statistique*, n° 212, juillet-août 1988.

Michel GLAUDE, Jean-Pierre JAROUSSE, "L'horizon des jeunes salariés dans leur entreprise", *Économie et Statistique*, n° 211, juin 1988.

Francis KRAMARZ, Stéfan LOLLIVIER, "Les difficultés de recrutement s'accentuent à la fin de 1989", *Économie et Statistique*, n° 234, juillet-août 1990.

Nicolas HERPIN, "La famille à l'épreuve du chômage", *Économie et Statistique*, n° 235, septembre 1990.

Eric MAURIN, "Les étrangers : une main-d'oeuvre à part ?" *Économie et Statistique*, n° 242, avril 1991.

Mireille MOUTARDIER, "Les conditions de vie des étrangers se sont améliorées depuis dix ans", *Économie et Statistique*, n° 242, avril 1991.

PROBIT

Stéfan LOLLIVIER, "Revenu offert, prétentions salariales et activité des femmes mariées", *Économie et Statistique*, n° 167, juin 84.

LOGIT (modèle polytomique non ordonné)

Luc ARRONDEL, André MASSON, "Hypothèse du cycle de vie, diversification et composition du patrimoine : France 1986", *Annales d'Économie et de Statistique*, n° 17, janvier-mars 1990.

Pour des références théoriques complètes :

Christian GOURIEROUX, *Économétrie des variables qualitatives*, éditions Economica, 1989, 2e édition.

Alan AGRESTI, *Categorical data Analysis*, John Wiley & Sons, 1990.

et un survey un peu ancien :

T. AMEMIYA, "Qualitative response models : a survey", *Journal of economic litterature*, vol. XIX, pp 1483-1536, décembre 1981.

Sur la partie du document de travail F9110 "Quelques problèmes économétriques souvent ignorés" :

Michael LECHNER, "Testing logit models in practice", document de travail, université d'Heidelberg (se le procurer auprès des auteurs de cette note).

H. WHITE, "Maximum Likelihood Estimation of Misspecified Models", *Econometrica* 50 : 1-25, 1982.

A. CHESHER, "Testing for Neglected Heterogeneity", *Econometrica* 52 : 865-872, 1984.

T. LANCASTER, "The Covariance Matrix of the Information Matrix Test", *Econometrica* 52 : 1051-1053, 1984.

C. ORME, "The Calculation of the Information Matrix Test for Binary Data Models", *The Manchester School* 56 : 370-376.

Pour la procédure LOGISTIC :

SAS/STAT User's guide, Volume 2, GLM-VARCOMP, Version 6, fourth edition.

(il existe une traduction très provisoire de PROC LOGISTIC à la division Études Sociales).