UNE ÉTUDE COMPARATIVE DES MÉTHODES DE DISCRIMINATION ET DE RÉGRESSION LOGISTIQUE

Olivier Sautory, Chang Way Sébastien Vong

Comparaison entre l'analyse discriminante linéaire et la régression logistique

Les problèmes traités

Un premier problème est l'étude de la relation entre une variable qualitative Y "expliquée" et une ou plusieurs variables "explicatives" $X^1 \ldots X^p$: il s'agit d'un modèle "causal" (selon la terminologie de Mc Fadden [5]), ou "conditionnel". Un tel modèle sera spécifié par la loi conditionnelle de Y sachant $\underline{X} = X^1 \ldots X^p$. La modélisation de cette relation entre Y et $\underline{X}$ par la fonction cumulative d'une distribution logistique caractérise la méthode de la **régression logistique**.

Un deuxième problème est celui de la discrimination entre deux ou plusieurs classes (chaque classe étant définie par une modalité d'une variable qualitative Y), fondée sur l'observation d'une ou plusieurs variables $X^1 \ldots X^p$. Il s'agit d'un modèle "conjoint" (toujours selon Mc Fadden), sans idée de causalité, qui est spécifié par la loi du couple $(Y, \underline{X})$. La modélisation des lois conditionnelles de $\underline{X}$ sachant Y par des lois normales de même matrice de variance-covariance, et de la loi de Y par une loi binomiale, caractérise la méthode d'**analyse discriminante linéaire.**

Ces deux approches conduisent à un problème commun : celui du classement d'une observation dans l'une des classes (i.e. l'une des modalités de Y) connaissant la valeur $\underline{X} = \underline{x}$. Il faut toutefois noter que ce problème de classement n'est pas à la base le problème traité par la régression logistique.

Dans toute la suite, on supposera, pour simplifier l'exposé, que la variable Y n'a que deux modalités.

Les méthodologies : modèle et estimation

Soit Y une variable dichotomique, prenant les modalités 0 et 1: elle définit deux groupes d'individus dans la population étudiée, le groupe 0 et le groupe 1. Soit $\underline{X}$ un vecteur de p variables quantitatives ($X^1 \ldots X^j \ldots X^p$)

On suppose que l'on a un échantillon de n observations indépendantes $(\underline{x}_i, y_i)$ du couple $(\underline{X}, Y)$ (avec $\underline{x}_i = (x_i^1 \ldots x_i^j \ldots x_i^p)$.

L'analyse discriminante linéaire (A.D.L.)

On suppose que les lois conditionnelles de $\underline{X}$ sachant Y sont des lois normales, de moyennes $\underline{\mu}_1$ et $\underline{\mu}_o$, et de même matrice de variance-covariance Σ :

$$f(\underline{x}/y=1) = f_1(\underline{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_1)' \Sigma^{-1} (\underline{x}-\underline{\mu}_1)\right]$$

$$f(\underline{x}/y=0) = f_0(\underline{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_0)' \Sigma^{-1} (\underline{x}-\underline{\mu}_0)\right]$$

Par ailleurs, Y suit une loi binomiale B (1,p) :

$$P(Y=1) = p, \; P(Y=0) = 1-p.$$

La densité du couple $(\underline{x}, y)$ est [1]:

$$f(\underline{x},y) = f(y) f(\underline{x}/y)$$
$$= p^y (1-p)^{1-y} f_1(\underline{x})^y f_0(\underline{x})^{1-y}$$

La log-vraisemblance du modèle à n observations est égale à :

$$\log L = \sum_{i=1}^{n} [y_i \log p + (1-y_i) \log (1-p)]$$

$$+ \sum_{i=1}^{n} y_i [-\frac{p}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma - \frac{1}{2} (\underline{x}_i - \underline{\mu}_1)' \Sigma^{-1} (\underline{x}_i - \underline{\mu}_1)]$$

$$+ \sum_{i=1}^{n} (1-y_i) [-\frac{p}{2} \log 2\pi - \frac{1}{2} \log \det \Sigma - \frac{1}{2} (\underline{x}_i - \underline{\mu}_0)' \Sigma^{-1} (\underline{x}_i - \underline{\mu}_0)]$$

On estime les paramètres p, $\underline{\mu}_0$, $\underline{\mu}_1$, et Σ par la méthode du maximum de vraisemblance ; on obtient :

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

i.e. la proportion d'observations du groupe 1 dans l'échantillon

$$\hat{\underline{\mu}}_1 = \frac{\sum_{i=1}^{n} y_i \underline{x}_i}{\sum_{i=1}^{n} y_i} = \bar{\underline{x}}_1 \, , \quad \hat{\underline{\mu}}_0 = \frac{\sum_{i=1}^{n} (1-y_i) \underline{x}_i}{\sum_{i=1}^{n} (1-y_i)} = \bar{\underline{x}}_0$$

1 f est une notation signifiant "densité", quelle que soit la densité dont il s'agit.

i.e. les moyennes empiriques du vecteur $\underline{X}$ respectivement dans les groupes 1 et 0

$$\hat{\Sigma} = \frac{1}{n}\left[\sum_{i=1}^{n} y_i (\underline{x}_i - \underline{\bar{x}}_1)(\underline{x}_i - \underline{\bar{x}}_1)' + \sum_{i=1}^{n} (1 - y_i)(\underline{x}_i - \underline{\bar{x}}_0)(\underline{x}_i - \underline{\bar{x}}_0)' \right]$$

i.e. la matrice de variance-covariance "intragroupe" empirique calculée sur l'échantillon.

La régression logistique

On modélise directement les probabilités conditionnelles de la façon suivante :

$$P(Y = 1/\underline{X} = \underline{x}) = \frac{1}{1 + \exp(-\underline{x}'\underline{\beta})}$$

$$P(Y = 0/\underline{X} = \underline{x}) = \frac{\exp(-\underline{x}'\underline{\beta})}{1 + \exp(-\underline{x}'\underline{\beta})}$$

où β est un paramètre de dimension p.

Aucune hypothèse n'est faite sur la loi du vecteur $\underline{X}$, qui peut d'ailleurs ne pas en avoir (cas de séries temporelles, où X est le temps, de cofacteurs contrôlés...).

Remarque : dans le cas du modèle logit "avec terme constant", la variable constante égale à 1 figure parmi les variables explicatives ; on écrira plutôt :

$$P(Y = 1/\underline{X} = \underline{x}) = \frac{1}{1 + \exp(-\underline{\beta}_0 - \underline{x}'\underline{\beta}_1)}$$

$\underline{x}$ désignant alors le vecteur des "vraies" variables explicatives.

La densité conditionnelle de Y sachant $\underline{X}$ est :

$$f(y / \underline{X} = \underline{x}) = \left[\frac{1}{1 + \exp(-\underline{x}'\underline{\beta})}\right]^{y} \left[\frac{\exp(-\underline{x}'\underline{\beta})}{1 + \exp(-\underline{x}'\underline{\beta})}\right]^{1-y}$$

La log-vraisemblance du modèle à n observations est égale à :

$$\log L = \sum_{i=1}^{n} \left[y_i \log \frac{1}{1+\exp(-x_i' \underline{\beta})} + (1-y_i) \log \frac{\exp -(x_i' \underline{\beta})}{1+\exp(-x_i' \underline{\beta})} \right]$$

$$= \sum_{i=1}^{n} [-\log(1+\exp -(\underline{x}_i' \underline{\beta})) - (1-y_i)\, \underline{x}_i' \underline{\beta}]$$

On estime le paramètre $\underline{\beta}$ par la méthode du maximum de vraisemblance (conditionnel) ; contrairement à l'analyse discriminante linéaire, on n'obtient pas de formule explicite pour l'estimateur du maximum de vraisemblance, qui est calculé en utilisant un algorithme itératif. On note $\tilde{\underline{\beta}}$ l'estimateur obtenu, $V_{as}(\tilde{\underline{\beta}})$ sa matrice de variance-covariance asymptotique, et $\tilde{V}$ un estimateur de cette matrice [8]

L'ADL est un cas particulier de la régression logistique

En analyse discriminante linéaire, la probabilité a posteriori d'appartenance au groupe 1, ou probabilité conditionnelle de Y = 1 sachant $\underline{X} = \underline{x}$, est égale à :

$$P(Y=1/\underline{X}=\underline{x}) = \frac{p f_1(\underline{x})}{p f_1(\underline{x}) + (1-p) f_0(\underline{x})}$$

$$= \frac{1}{1+\frac{1-p}{p}\frac{f_0(\underline{x})}{f_1(\underline{x})}}$$

$$= \frac{1}{1+\frac{1-p}{p}\exp -\frac{1}{2}[(\underline{x}-\underline{\mu}_0)' \Sigma^{-1}(\underline{x}-\underline{\mu}_0) - (\underline{x}-\underline{\mu}_1)' \Sigma^{-1}(\underline{x}-\underline{\mu}_1)]}$$

$$= \frac{1}{1+\exp -[(\underline{\mu}_1-\underline{\mu}_0)' \Sigma^{-1} \underline{x} - \log\frac{1-p}{p} + \frac{1}{2}\underline{\mu}_0{}' \Sigma^{-1} \underline{\mu}_0 - \frac{1}{2}\underline{\mu}_1{}' \Sigma^{-1} \underline{\mu}_1)]}$$

$$= \frac{1}{1+\exp -[S(\underline{x})]}$$

où $S(\underline{x})$ est le "score", ou statistique d'Anderson, déduit de la fonction de Fisher, qui permet l'affectation d'une observation à l'un des deux groupes :

- on affecte $\underline{x}$ au groupe 1 si $S(\underline{x}) > 0$

- on affecte $\underline{x}$ au groupe 0 si $S(\underline{x}) < 0$

La probabilité a posteriori apparaît ainsi comme une fonction logistique du score.

Le modèle d'analyse discriminante linéaire est donc un cas particulier du modèle logistique ; il s'agit d'un modèle avec terme constant de la forme :

$$P(Y = 1/\underline{X} = \underline{x}) = \frac{1}{1 + \exp(-\beta_0 - \underline{x}' \underline{\beta}_1)}$$

avec :

$$\beta_0 = -\log \frac{1-p}{p} + \frac{1}{2} \underline{\mu}_0' \Sigma^{-1} \underline{\mu}_0 - \frac{1}{2} \underline{\mu}_1' \Sigma^{-1} \underline{\mu}_1$$

$$\underline{\beta}_1 = \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_0)$$

Les estimateurs du maximum de vraisemblance de β_0 et $\underline{\beta}_1$ sont, en vertu du principe d'invariance fonctionnelle :

$$\hat{\beta}_0 = -\log \frac{1-\hat{p}}{\hat{p}} + \frac{1}{2} \hat{\underline{\mu}}_0' \hat{\Sigma}^{-1} \hat{\underline{\mu}}_0 - \frac{1}{2} \hat{\underline{\mu}}_1' \hat{\Sigma}^{-1} \hat{\underline{\mu}}_1$$

$$\hat{\underline{\beta}}_1 = \hat{\Sigma}^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_0)$$

Quelles sont les propriétés respectives de $\hat{\underline{\beta}}$ $(= (\hat{\beta}_0, \hat{\underline{\beta}}_1))$, obtenu par l'analyse discriminante linéaire, et de $\tilde{\underline{\beta}}$, obtenu par la régression logistique ?

- si le modèle de l'analyse discriminante linéaire est vrai (et donc celui de la régression logistique l'est également) : $\hat{\underline{\beta}}$ est convergent et asymptotiquement efficace (puisque c'est l'estimateur du maximum de vraisemblance), alors que $\tilde{\underline{\beta}}$, s'il est convergent, est moins précis que $\hat{\underline{\beta}}$ (il ne prend pas en compte toute l'information du modèle). Les matrices de variance-covariance asymptotiques de $\hat{\underline{\beta}}$ et $\tilde{\underline{\beta}}$ vérifient :

$$V_{as}(\hat{\underline{\beta}}) << V_{as}(\tilde{\underline{\beta}})$$

- si le modèle de l'analyse discriminante linéaire n'est pas correct, alors que celui de la régression logistique l'est, $\hat{\underline{\beta}}$ n'est plus convergent, alors que $\tilde{\underline{\beta}}$ l'est, et est asymptotiquement efficace.

Quelques études "classiques"

Efron [2]

Efron a comparé les pouvoirs de discrimination respectifs de l'analyse discriminante linéaire et de la régression logistique lorsque l'hypothèse de normalité est vérifiée. Son étude est fondée sur le calcul des taux d'erreur (ou probabilités de mauvais classement) obtenus avec chacune des méthodes, en utilisant les distributions asymptotiques de $\hat{\underline{\beta}}$ et $\tilde{\underline{\beta}}$: le rapport de ces taux d'erreur asymptotiques, qu'il appelle efficacité relative asymptotique ERA (discriminante/logit), est une fonction décroissante de la distance de Mahalanobis $D^2 = (\underline{\mu}_1 - \underline{\mu}_0)' V^{-1} (\underline{\mu}_1 - \underline{\mu}_0)$ entre les deux populations normales. Le tableau suivant correspond au cas où $p = 1\text{-}p = 1/2$ (qui est le cas le plus favorable à la régression logistique).

D	0	0.5	1	1.5	2	2.5	3	3.5	4
ERA	1.000	1.000	0.995	0.968	0.899	0.786	0.641	0.486	0.343

Ces résultats ne sont bien sûr pas surprenants, puisque l'on sait que, sous l'hypothèse de normalité, $\hat{\underline{\beta}}$ est asymptotiquement efficace.

Efron a noté que la modélisation logistique de la vraisemblance conditionnelle est valable dès que les lois conditionnelles $f_1(\underline{x})$ et $f_2(\underline{x})$ appartiennent à la famille exponentielle, i.e. sont de la forme :

$$f_1(\underline{x}) = g(\underline{\theta}_1, \underline{\eta})\, h(\underline{x}, \underline{\eta}) \exp(\underline{\theta}_1' \underline{x})$$

$$f_0(\underline{x}) = g(\underline{\theta}_0, \underline{\eta})\, h(\underline{x}, \underline{\eta}) \exp(\underline{\theta}_0' \underline{x})$$

où $\underline{\eta}$ est un paramètre nuisible, comme Σ dans le cas où les lois f_1 et f_0 sont normales (cas particulier de la formulation générale précédente).

On a alors en effet :

$$P(Y=1/\underline{X}=\underline{x}) = \frac{1}{1+\exp-(\alpha+\underline{\beta}'\,\underline{x})}$$

avec :

$$\underline{\beta} = \underline{\theta}_1 - \underline{\theta}_0$$

$$\alpha = \log\left[\frac{g(\underline{\theta}_1, \underline{\eta})}{g(\underline{\theta}_0, \underline{\eta})}\right] - \log\frac{\pi_0}{\pi_1}$$

Press et Wilson [7]

Press et Wilson préconisent l'usage de la régression logistique de préférence à l'analyse discriminante linéaire en raison de sa plus grande robustesse lorsque l'hypothèse de normalité n'est pas vérifiée. Ils illustrent leur propos à l'aide de deux études empiriques : dans chacune d'elles, les variables explicatives sont un mélange de variables quantitatives et de variables dichotomiques (associées aux modalités de variables qualitatives). En partageant, dans chaque exemple, les observations en deux groupes (échantillon de base, qui sert à calculer les estimateurs, et échantillon-test, qui permet d'estimer les probabilités de "bon classement"), les auteurs ont constaté la supériorité de la régression logistique "but not by a large amount".

Amemyia et Powell [1]

Cette relativement bonne performance de l'analyse discriminante linéaire en présence de variables binaires (variables 0-1), donc lorsque l'hypothèse de normalité est singulièrement violée, a directement motivé l'étude de Amemyia et Powell : ils ont étudié le cas où les variables explicatives sont indépendantes et binaires, dans un cadre asymptotique.

Par des évaluations numériques des résultats asymptotiques auxquels ils parviennent, ils ont montré que l'analyse discriminante linéaire "does quite well" en termes de probabilités de bon classement, et "does mostly well" en termes de précision des estimations (mesurée par l'erreur quadratique moyenne pour l'analyse discriminante).

Les auteurs expliquent l'excellent résultat de l'analyse discriminante linéaire en termes de classement des observations, en remarquant que, lorsque les variables explicatives sont discrètes, une infinité de fonctions linéaires discriminantes, avec des coefficients pouvant varier largement, peuvent conduire au même classement. Des "imprécisions" sur les estimations peuvent être sans effet sur la (bonne) qualité de la règle d'affectation.

Test de l'analyse discriminante linéaire contre le modèle logit

Le choix entre analyse discriminante linéaire et régression logistique pourrait reposer sur l'acceptation, ou le rejet, de l'hypothèse de normalité de la loi de $\underline{X}$ sachant Y. On pourrait utiliser les tests de normalité, classiques dans le cas d'une variable :

- test de Shapiro-Wilks ;
- test de Kolmogorov ;
- test fondé sur le coefficient d'asymétrie ("skewness") ;
- test fondé sur le coefficient d'aplatissement ("kurtosis") etc.

et un peu moins classiques dans le cas multivarié (SAS ne propose aucun test de ce type).

Mais dans la mesure où l'on accepte l'idée que l'on utilise soit l'analyse discriminante linéaire, soit la régression logistique, pourquoi ne pas tester directement l'une contre l'autre : c'est la démarche proposée par Lo [6] , fondée sur un test de spécification d'Hausman [4].

Les deux hypothèses du test peuvent être formulées ainsi :

- $H_0 : f(\underline{X} / Y = y)$ est une loi normale de moyenne $\underline{\mu}_y$ et de matrice de variance-covariance Σ
- $H_1 : f(\underline{X} / Y = y)$ est une loi de la famille exponentielle.

Le test d'Hausman consiste à former la statistique :

$$J = n\,(\tilde{\underline{\beta}} - \hat{\underline{\beta}})'\,(\tilde{V} - \hat{V})^{-1}\,(\tilde{\underline{\beta}} - \hat{\underline{\beta}})$$

avec :

$\tilde{V}$ = estimation de $V_{as}(\tilde{\underline{\beta}})$

$\hat{V}$ = estimation de $V_{as}(\hat{\underline{\beta}})$

Lo propose une estimation de $V_{as}(\hat{\underline{\beta}})$, qui n'est malheureusement pas aisément calculable dans SAS.

Sous H_0 , J converge en loi vers un chi-deux à p degrés de liberté (p = nombre de variables explicatives, y compris la constante).

On refuse donc H_0 si J est supérieur au quantile d'ordre 0.95 (dans le cas d'un test de niveau 5 %) de la loi du chi-deux à p degrés de liberté.

Autres éléments de comparaison entre l'A.D.L. et la régression logistique

1. La régression logistique demande l'estimation d'un nombre de paramètres inférieur à celui de l'analyse discriminante linéaire (les moyennes de chaque groupe et la matrice de variance-covariance commune) : cette estimation peut donc s'avérer plus précise, en particulier dans le cas de faibles échantillons.

2. Les temps de calcul sont généralement plus longs (de l'ordre de 50 %) pour la régression logistique, en raison de la méthode d'estimation des paramètres (algorithme itératif).

3. L'analyse discriminante linéaire et la régression logistique se comportent différemment lorsqu'il y a séparation totale des deux groupes : l'analyse discriminante fait "comme si de rien n'était", alors que la régression logistique refuse de calculer les estimateurs des paramètres (car elle ne le peut pas...). Ce comportement de la régression logistique permet donc d'alerter l'utilisateur vigilant sur le caractère particulier de ses données.

Les données suivantes, extraites de Press-Wilson [7], illustrent cette situation (il n'y a ici qu'une seule variable explicative).

obs i	1	2	3	4	5	6	7	8
x (i)	- 4	- 3	- 2	- 1	1	2	3	4
y (i)	0	0	0	0	1	1	1	1

Pour la régression logistique, l'une des deux équations conduisant à l'estimation des paramètres θ_0 et θ_1 est :

$$\sum_{i=1}^{8} x_i \frac{1}{1+\exp(-\theta_0-\theta x_i)} = \sum_{i=1}^{8} x_i y_i = 10$$

Elle n'a pas de solution.

En revanche, l'analyse discriminante linéaire conduit à l'estimation suivante de θ_1 :

$$\hat{\theta}_1 = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}^2} = 4$$

BIBLIOGRAPHIE

[1] AMEMYIA, T., & POWELL, J. L. (1983), A comparison of the logit model and normal discriminant analysis when the independant variables are binary, In: *Studies in Econometrics, Time series and Multivariate Statistics*, Ed. S. Karlin, T. Amemyia and L. A. Goodman, pp. 3-30. New York: Academic Press.

[2] EFRON, B. (1975), The efficiency of logistic regression compared to discriminant analysis J., *Am. Statist. Assoc.* 70, pp. 892-898.

[3] GOURIEROUX, C. (1991), Les scores, *Polycopié ENSAE.*

[4] HAUSMAN, J. (1978), Specification tests in econometrics, *Econometrica* 46, pp. 1251-1271.

[5] Mc FADDEN D. (1976), A comment on discriminant analysis 'versus' logit analysis. *Annals of economic and social measurement* 5, pp. 511-523.

[6] LO, A. W. (1986), Logit versus discriminant analysis, A specification test and application to corporate bankruptcies, *Journal of Econometrics* 31, pp. 151-178.

[7] PRESS, S. J. & WILSON, S. (1978), Choosing between logistic regression and discriminant analysis, *J. Am. Statist. Assoc.* 73, pp. 699-705.

[8] MARPSAT, M. & TROGNON, A., (1992). L'usage des modèles logit : présentation générale, In *Journées de méthodologie statistique de l'Insee 1992.*

Une étude empirique

Les données

Les données utilisées pour notre étude proviennent de l'enquête" Budget de famille" réalisée par l'Insee en 1989, auprès de 9 038 ménages.

La variable qualitative Y "expliquée" est la variable "possession d'une carte de crédit dans le ménage", notée CARCRE : elle vaut 1 si le ménage possède une carte, 2 sinon. Dans l'échantillon, 52 % des ménages détiennent une carte de crédit.

La richesse du questionnaire fournissait *a priori* un nombre élevé de variables "explicatives" potentielles $X^1 \ldots X^p$. Nous avons opéré une première sélection d'une vingtaine de variables qui nous ont semblé pertinentes par rapport au phénomène étudié. Nous avons ensuite sélectionné les variables les plus liées à la variable CARCRE, en utilisant pour cela le critère de V de Cramer (dérivé de la statistique du chi-deux d'un tableau de contingence), et en évitant de choisir des variables qui soient trop liées entre elles. Nous avons ainsi retenu 7 variables : la catégorie socioprofessionnelle de la personne de référence, son âge (par tranche), le statut d'occupation du logement, le nombre de personnes dans le ménage, la catégorie de commune, le revenu mensuel total du ménage, et le montant total des crédits payés par le ménage au cours des 12 mois précédents. Toutes ces variables sont qualitatives, sauf la dernière, qui a toutefois été rendue qualitative par découpage en classes. La liste des modalités de ces variables figure à l'*annexe 1*.

Les méthodes utilisées

L'analyse discriminante

Les variables explicatives étant qualitatives, nous ne pouvons pas utiliser directement - en principe tout au moins... - les procédures d'analyse discriminante du logiciel SAS, qui sont adaptées aux variables quantitatives. Nous avons donc utilisé la procédure suivante, proposée par Saporta, que nous appellerons DISQUAL :

- on réalise dans un premier temps une analyse des correspondances multiples sur l'ensemble des variables qualitatives $X^1 \ldots X^p$, i.e. une analyse factorielle des correspondances sur le tableau formé par les variables indicatrices (variables 0-1) associées aux modalités de $X^1 \ldots X^p$.

- on récupère un certain nombre de variables factorielles issues de cette analyse (i.e. les coordonnées des observations sur les axes factoriels) $c^1 \ldots c^q$: ces variables quantitatives sont utilisées pour réaliser l'analyse discriminante linéaire (qui suppose la normalité des variables dans chaque groupe, avec des matrices de variance-covariance égales)

- la fonction discriminante de Fisher obtenue est une combinaison linéaire des variables $c^1 \ldots c^q$, qui sont elles-mêmes combinaisons linéaires des variables indicatrices en entrée de l'analyse des correspondances : la fonction de Fisher est donc elle-même une combinaison linéaire de ces indicatrices des modalités. Les coefficients de cette combinaison sont appelés les "scores" des modalités. La valeur de la fonction de Fisher pour un individu, qui permet son classement dans l'un ou l'autre groupe, s'obtient donc en faisant la somme des scores des modalités que prend l'individu.

Dans la pratique, se pose le problème du choix des axes factoriels pour lesquels on récupère les coordonnées : en général, on utilise les premiers axes, en se limitant à ceux qui séparent bien les deux groupes.

Pour utiliser cette méthode à l'aide de SAS, nous avons utilisé une "macro", appelée DISQUAL, qui enchaîne l'analyse des correspondances multiples (procédure CORRESP) et l'analyse discriminante (procédure DISCRIM). Cette macro permet le choix des facteurs de l'ACM, calcule pour chacun d'eux le "rapport de corrélation" (égal au rapport : variance intergroupe / variance totale) mesurant le pouvoir séparateur entre les groupes. Elle produit les résultats (succincts) de l'analyse discriminante, un tableau donnant les scores des modalités des variables qualitatives, et un tableau donnant pour chaque individu son score, ses probabilités a posteriori et son groupe d'affectation.

La régression logistique

En régression logistique, lorsque les variables explicatives sont qualitatives, on utilise les variables indicatrices associées aux modalités des variables : ainsi, la variable catégorie socioprofessionnelle (CS) à 8 modalités est remplacée par 8 variables indicatrices (CS_1 , CS_2 ,... , CS_8). Mais un problème de colinéarité apparaît : la somme des m indicatrices d'une variable à m modalités vaut toujours 1. Pour remédier à ce problème, on élimine une modalité pour chaque variable : ceci revient à donner la valeur 0 au coefficient associé à cette modalité, que l'on appelle "situation de référence", par rapport à laquelle on mesure les écarts. Le choix de cette situation de référence n'influe pas sur les résultats, à une translation des coefficients près. La pratique veut que l'on choisisse en général la modalité la plus fréquente.

Pour mettre en oeuvre la méthode de la régression logistique, nous avons utilisé la procédure LOGISTIC de SAS.

Première analyse

Nous comparons les résultats fournis par les deux méthodes dans une première analyse, avec les variables explicatives indiquées précédemment ; les 10 premiers axes de l'ACM ont été sélectionnés pour l'analyse discriminante. Les résultats sont donnés à l'annexe 2. On constate que la régression logistique donne un taux d'erreur de classement de 28.3 %, légèrement inférieur à celui obtenu avec l'analyse discriminante (29.8 %).

Classement des variables explicatives

On peut classer les variables explicatives selon un "pouvoir explicatif" décroissant, à l'aide d'indicateurs adéquats. Les indicateurs que nous avons utilisés ici sont les suivants :

- en analyse discriminante : nous avons calculé, pour chaque variable qualitative à m modalités, l'écart entre les deux scores extrêmes parmi les m scores donnés par l'analyse ;

- en régression logistique : nous avons calculé, pour chaque variable qualitative à m modalités, l'écart entre les deux coefficients extrêmes parmi les m coefficients donnés par la régression (parmi ces coefficients extrêmes peut éventuellement figurer la valeur 0 associée à la situation de référence, dans le cas où les autres coefficients seraient tous positifs ou tous négatifs).

Les résultats figurent dans le tableau ci-dessous :

Variables expliquées	Analyse discrimante	Régression logistique
CS	1.06	1.72
TRAGECH	1.02	2.27
REVENS	0.99	2.26
NBPERS	0.69	0.59
C	0.66	0.25
DETTE	0.54	0.87
STALOG	0.53	0.32

Les hiérarchies issues des deux méthodes ne sont pas exactement identiques, même s'il apparaît dans les deux cas que trois variables semblent prédominantes pour expliquer la possession d'une carte de crédit : la catégorie socioprofessionnelle, l'âge et le revenu.

Étude par variable

Pour chaque variable, on peut classer les modalités selon leur effet sur la possession d'une carte de crédit, en utilisant les scores de l'analyse discriminante, et les coefficients de la régression logistique : les scores et les coefficients positifs correspondent à des modalités "favorables" à la possession de carte de crédit. Le tableau de l'annexe 3 permet de comparer les classements donnés par les deux méthodes.

La catégorie socioprofessionnelle (CS)

En logistique, la modalité de référence est celle des ouvriers. Par rapport à ceux-ci, les cadres supérieurs, les professions intermédiaires et les employés ont plus de chance de détenir une carte de crédit. Ce résultat se retrouve aussi en analyse discriminante. En revanche, les agriculteurs et les retraités ne possèdent pas la même chance d'avoir une carte de crédit en analyse discriminante et en logistique.

L'âge de la personne de référence (TRAGECH)

En logistique, les coefficients associés aux différentes tranches d'âge diminuent avec celles-ci : plus les ménages sont jeunes, plus il est probable qu'ils détiennent une carte de crédit. On observe également ce phénomène en analyse discriminante, quoique moins systématique.

Le statut d'occupation du logement (STATLOG)

Les deux classements sont ici assez différents, puisque les ménages accédant à la propriété sont placés en première position par l'analyse discriminante, et en dernière par la régression logistique ; notons toutefois que les effets de cette variable sont parmi les plus faibles, deux des trois coefficients de la régression logistique n'étant d'ailleurs pas significatifs (statistique de Wald trop faible).

La catégorie de commune (C)

Les deux classements sont sensiblement différents, mais on observe dans les grandes unités urbaines la propension à détenir une carte de crédit plus élevée que dans les petites communes ou agglomérations.

Le nombre de personnes du ménage (NBPERS)

On constate ici une divergence importante entre les méthodes : en logistique, les coefficients décroissent avec le nombre de personnes du ménage, alors qu'en analyse discriminante les ménages d'une personne ont le score le plus faible, et les ménages de 4 personnes le score le plus fort. C'est d'ailleurs de ce que l'on observe avec les taux de possession "bruts" dans l'échantillon, qui valent respectivement, pour ces deux catégories 36 % et 68 %. Ceci voudrait-il dire que le régression logistique opère plus "toutes choses égales par ailleurs" que ne le fait l'analyse discriminante ?

Le revenu du ménage (REVENS)

Cette variable joue de la même façon pour les deux méthodes : plus on est riche, plus on a de chance de posséder une carte de crédit...

La dette annuelle du ménage (DETTE)

Les deux classements coïncident pour cette variable également : les ménages fortement endettés ont une probabilité élevée d'avoir une carte de crédit.

Utilisation des intervalles de confiance

Les coefficients estimés par la régression logistique, aussi bien que les scores calculés par l'analyse discriminante, sont bien sûr affectés d'une certaine précision. La procédure LOGISTIC permet d'évaluer cette précision, car elle calcule, pour chaque coefficient estimé β_i, une estimation σ_i de son écart-type (asymptotique) : on peut en déduire des intervalles de confiance de la forme :

$$[\beta_i - 2\sigma_i, \beta_i + 2\sigma_i]$$

Malheureusement, la procédure DISCRIM ne donne pas le même genre d'informations. Il nous a paru toutefois intéressant de calculer les intervalles de confiance des coefficients estimés par la régression logistique. À titre d'exemple, le tableau suivant concerne la variable catégorie socioprofessionnelle (les modalités sont classées par coefficient estimé décroissant). On constate que les intervalles associés aux modalités d'une même variable se chevauchent largement, ce qui est peut-être de nature à nuancer certaines des divergences constatées précédemment (cas des modalités 2 et 7 par exemple).

modalités	régression logistique			analyse discrimante
3	0.53	-	0.99	0.67
4	0.47	-	0.81	0.51
5	0.24	-	0.59	0.30
1	0.05	-	0.62	- 0.29
7	- 0.05	-	0.39	- 0.39
2	- 0.20	-	0.27	- 0.20
8	- 0.25	-	0.28	- 0.29
6		0		- 0.09

Remarquons par ailleurs la longueur de ces intervalles de confiance, malgré la taille importante de l'échantillon (8990 ménages).

Deuxième analyse

Pour tenter de comprendre d'où proviennent les divergences entre les résultats fournis par les deux méthodes statistiques, nous avons introduit les variables exogènes les unes après les autres dans le modèle. La divergence entre les deux méthodes apparaît lorsque les variables NBPERS, REVENS sont introduites. Parmi les variables explicatives, NBPERS et REVENS sont les plus corrélées entre elles ; on pouvait donc se demander si les différences constatées ne proviennent pas d'une prise en compte distincte des interactions. Nous avons donc construit une variable croisant NBPERS et REVENS, notée NBPREV, dont les modalités sont indiquées ci-dessous :

Modalités : (revenu mensuel)

01 : ménage d'une personne, revenu moins de 7 800 F ;

02 : ménage d'une personne, revenu de 7 800 F à moins de 18 000 F ;

03 : ménage d'une personne, revenu de 18 000 F et plus ;

04 : ménage de deux personnes, revenu moins de 7 800 F ;

05 : ménage de deux personnes, revenu de 7 800 F à moins de 18 000 F

06 : ménage de deux personnes, revenu de 18 000 F et plus ;

07 : ménage de trois personnes, revenu moins de 7 800 F ;

08 : ménage de trois personnes, revenu de 7 800 F à moins de 18 000 F ;

09 : ménage de trois personnes, revenu de 18 000 F et plus ;

10 : ménages de quatre personnes, revenu moins de 7 800 F ;

11 : ménage de quatre personnes revenu de 7 800 F à moins de 18 000 F ;

12 : ménage de quatre personnes, revenu de 18 000 F et plus ;

13 : ménage de cinq personnes et plus, revenu moins de 7 800 F ;

14 : ménage de cinq personnes et plus, revenu de 7 800 F à moins de 18 000 F ;

15 : ménage de cinq personnes et plus, revenu de 18 000 F et plus.

Nous avons ensuite réalisé une deuxième analyse, avec les mêmes variables que précédemment, exceptées NBPERS et REVENS remplacées par NBPREV : les divergences demeurent. Pour la variable NBPREV, la régression logistique classe d'abord par revenu, puis par taille du ménage (de 1 à 5 sauf pour les faibles revenus), alors que l'analyse discriminante mêle un peu plus les deux variables, même si l'effet revenu est largement prédominant.

en logistique :

NBPREV (Nombre de personnes du ménage X Revenu mensuel) :

Modalités

03 : ménage d'une personne, revenu de 18 000 F et plus ;

06 : ménage de deux personnes, revenu de 18 000 F et plus ;

09 : ménage de trois personnes, revenu de 18 000 F et plus ;

12 : ménage de quatre personnes, revenu de 18 000 F et plus ;

15 : ménage de cinq personnes et plus, revenu de 18 000 F et plus ;

02 : ménage d'une personne, revenu de 7 800 F à moins de 18 000 F ;

05 : ménage de deux personnes, revenu de 7 800 F à moins de 18 000 F ;

08 : ménage de trois personnes, revenu de 7 800 F à moins de 18 000 F ;

11 : ménage de quatre personnes, revenu de 7 800 F à moins de 18 000 F ;

14 : ménage de cinq personnes et plus, revenu de 7 800 F à moins de 18 000 F ;

10 : ménagede quatre personnes, revenu moins de 7 800 F ;

01 : ménage d'une personne, revenu moins de 7 800 F ;

04 : ménage de deux personnes, revenu moins de 7 800 F ;

07 : ménage de trois personnes, revenu moins de 7 800 F ;

13 : ménage de cinq personnes et plus, revenu moins de 7 800 F ;

L'analyse discriminantes :

NBPREV (nombre de personnes du ménage revenu mensuel) :

Modalités :

12 : ménage de quatre personnes, revenu de 18 000 F et plus ;

15 : ménage de cinq personnes et plus, revenu de 18 000 F et plus ;

06 : ménage de deux personnes, revenu de 18 000 F et plus ;

09 : ménage de trois personnes, revenu de 18 000 F et plus ;

02 : ménage d'une personne, revenu de 7 800 F à moins de 18 000 F ;

03 : ménage d'une personne, revenu de 18 000 F et plus ;

11 : ménage de quatre personnes, revenu de 7 800 F à moins de 18 000 F ;

08 : ménage de trois personnes, revenu de 7 800 F à moins de 18 000 F ;

05 : ménage de deux personnes, revenu de 7 800 F à moins de 18 000 F ;

07 : ménage de trois personnes, revenu moins de 7 800 F ;

13 : ménage de cinq personnes et plus, revenu moins de 7 800F ;

14 : ménage de cinq personnes et plus, revenu de 7 800 F à moins de 18 000 F ;

01 : ménage d'une personne, revenu moins de 7 800 F ;

04 : ménage de deux personnes, revenu moins de 7 800 F ;

10 : ménage de quatre personnes, revenu moins de 7 800 F ;

Analyses complémentaires

Une nouvelle variable expliquée

Dans cette analyse, la variable expliquée est la possession d'une assurance-vie. Les résultats sont comparables à ceux obtenus avec la variable possession d'une carte de crédit : taux d'erreur de classement valant 29,8 % en logistique et 30,4 % en analyse discriminante, certaines divergences, semblables en nombre et en ampleur à celles de l'analyse précédente, entre les classements des modalités au sein des variables explicatives, et des hiérarchies de variables quelque peu différentes, comme le montre le tableau suivant :

Variables expliquées	Analyse discriminante	Régression logistique
CS	0.90	1.19
TRAGECH	0.78	0.82
REVENS	0.64	1.09
NBPERS	0.52	0.41
STALOG	0.51	0.40
DETTE	0.38	0.28
C	0.16	0.25

Lorsque l'on prend tous les axes de l'ACM...

Nous avons réalisé une nouvelle analyse discriminante, avec les variables de la deuxième analyse, mais en sélectionnant cette fois tous les axes issus de l'analyse des correspondances multiples. Les résultats deviennent alors presque identiques à ceux de la régression logistique, tant en ce qui concerne la hiérarchie des variables que le classement des modalités. Les divergences entre les deux méthodes semblent donc provenir pour l'essentiel du fait qu'avec la méthode DISQUAL, on n'utilise, pour l'analyse discriminante, qu'une partie de l'information du tableau de départ dès que l'on ne sélectionne pas tous les axes issus de l'ACM. Nous proposons donc la conclusion provisoire suivante :

- si l'on désire utiliser toute l'information du tableau, régression logistique et analyse discriminante sont équivalentes ; la régression logistique est plus simple à mettre en oeuvre, et fournit des intervalles de confiance sur les coefficients ;

- si l'on pense qu'il vaut mieux commencer par éliminer le "bruit" qui se trouve dans les données, pour ne garder que l'"information significative", il est préférable d'utiliser la méthode DISQUAL.

ANNEXE 1

C (Catégorie socioprofessionnelle de la personne de référence) :

Modalité :

1 : Agriculteurs exploitants
2 : Artisans, commerçant et chefs d'entreprise
3 : Cadres et professions intellectuelles supérieures
4 : Professions intermédiaires
5 : Employés
6 : Ouvriers
7 : Retraités
8 : Autres personnes sans activité professionnelle

TRAGEC (Tranche d'âge de la personne de référence) :

Modalités :

1 : Moins de 25 ans
2 : De 25 à moins de 35 ans
3 : De 35 à moins de 45 ans
4 : De 45 à moins de 55 ans
5 : De 55 à moins de 65 ans
6 : De 65 à moins de 75 ans
7 : 75 ans et plus

NBPERS (Nombre de personnes du ménages) :

Modalités :

1 : 1 personne
2 : 2 personnes
3 : 3 personnes
4 : 4 personnes
5 : 5 personnes

STALOGN (Statut d'occupation du logement) :

Modalités :

1 : Propriétaire ou copropriétaire
2 : Accédant à la propriété
3 : Locataire, sous-locataire
4 : Logé gratuitement

C (Catégorie de commune) :

Modalités :

0 : Communes rurales
1 : Unités urbaines de moins de 20 000 habitants
2 : Unités urbaines de 20 000 à moins de 100 000 habitants

3 : Unités urbaines de 100 000 habitants et plus
4 : Complexe de l'agglomération parisienne (yc Paris)

REVENSN (Revenu mensuel par tranche de tous les ménages) :

Modalités :

1 : Moins de 5800 F
2 : De 5 800 F à moins de 7 800 F
3 : De 7 800 F à moins de 11 000 F
4 : De 11 000 F à moins de 18 000 F
5 : De 18 000 F à moins de 37 000 F
6 : 37 000 F et plus

DETTEN (Montant total des crédits payés par le ménage au cours des 12 derniers mois) :

Modalités :

1 : Pas de dette
2 : Ayant une dette de moins de 25 000 F
3 : 25 000 F et plus

Discriminant Analysis

8990 Observations	8989 DF Total
10 Variables	8988 DF Within Classes
2 Classes	1 DF Between Classes

Class Level Information

CARCRE	Output SAS Name	Frequency	Weight	Proportion	Prior Probability
1	_1	4661	4661	0.518465	0.518465
2	_2	4329	4329	0.481535	0.481535

Analyse discriminante linéaire sur les facteurs de l'ACM sélectionnés

Discriminant Analysis Classification Summary for Calibration Data: WORK._8_

Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D^2_j(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j) - 2 \ln PRIOR_j$$

Posterior Probability of Membership in each CARCRE:

$$Pr(j|X) = \exp(-.5\, D^2_j(X)) \,/\, SUM_k \exp(-.5\, D^2_k(X))$$

Number of Observations and Percent Classified into CARCRE:

From CARCRE	1	2	Total
1	3511	1150	4661
	75.33	24.67	100.00
2	1527	2802	4329
	35.27	64.73	100.00
Total	5038	3952	8990
Percent	56.04	43.96	100.00
Priors	0.5185	0.4815	

Error Count Estimates for CARCRE:

	1	2	Total
Rate	0.2467	0.3527	0.2978
Priors	0.5185	0.4815	

Analyse de la variance : dispersion de la variable CARCRE sur chacun des 10 axes de l'ACM sélectionnés

OBS	FACTEUR	R2	FISHER	DL1	DL2	PROB
1	DIM1	0.18174	1996.35	1	8988	0.00000
2	DIM2	0.00296	26.72	1	8988	0.00000
3	DIM3	0.00722	65.38	1	8988	0.00000
4	DIM4	0.01917	175.67	1	8988	0.00000
5	DIM5	0.00043	3.90	1	8988	0.04838
6	DIM6	0.00561	50.69	1	8988	0.00000
7	DIM7	0.00000	0.04	1	8988	0.84731
8	DIM8	0.00004	0.40	1	8988	0.52755
9	DIM9	0.00070	6.28	1	8988	0.01225
10	DIM10	0.00632	57.16	1	8988	0.00000

ESSAI DE LA MACRO DISQUAL : DISCRIMINATION SUR VARIABLES QUALITATIVES

Scores des modalités des variables qualitatives (coefficients de la différence entre la fonction de Fisher relative à la 1ère modalité de CARCRE (1) et la fonction de Fisher relative à la 2ème modalité de CARCRE (2))

OBS	MENA	SCORE	CONSTANT
1	CS1	-0.29245	0.095264
2	CS2	-0.20452	0.095264
3	CS3	0.67196	0.095264
4	CS4	0.51218	0.095264
5	CS5	0.29807	0.095264
6	CS6	-0.09315	0.095264
7	CS7	-0.39222	0.095264
8	CS8	-0.28962	0.095264
9	TRAGECH1	0.25273	0.095264
10	TRAGECH2	0.28476	0.095264
11	TRAGECH3	0.15759	0.095264
12	TRAGECH4	0.17317	0.095264
13	TRAGECH5	-0.12061	0.095264
14	TRAGECH6	-0.26565	0.095264
15	TRAGECH7	-0.73611	0.095264
16	STALOGN1	-0.27118	0.095264
17	STALOGN2	0.25810	0.095264
18	STALOGN3	0.04671	0.095264
19	STALOGN4	0.11137	0.095264
20	C0	-0.17895	0.095264
21	C1	-0.22326	0.095264
22	C2	-0.24156	0.095264
23	C3	0.22382	0.095264
24	C4	0.42269	0.095264
25	NBPERS1	-0.32751	0.095264
26	NBPERS2	-0.03357	0.095264
27	NBPERS3	0.18424	0.095264
28	NBPERS4	0.36022	0.095264
29	NBPERS5	-0.19363	0.095264
30	REVENSN1	-0.45535	0.095264
31	REVENSN2	-0.09113	0.095264
32	REVENSN3	-0.08298	0.095264
33	REVENSN4	0.15455	0.095264
34	REVENSN5	0.53438	0.095264
35	REVENSN6	0.48400	0.095264
36	DETTEN1	-0.25059	0.095264
37	DETTEN2	0.17358	0.095264
38	DETTEN3	0.28507	0.095264

ANNEXE 2 (suite et fin)

The SAS System

The LOGISTIC Procedure

Data Set: WORK.A
Response Variable: CARCRE 21000000 CARCRE POS=291
Response Levels: 2
Number of Observations: 8990
Link Function: Logit

Response Profile

Ordered Value	CARCRE	Count
1	1	4661
2	2	4329

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1.7227	0.2720	40.0843	0.0001	.
CS1	0.3344	0.1410	5.6218	0.0177	0.033639
CS2	0.0328	0.1179	0.0774	0.7809	0.003870
CS3	0.7639	0.1149	44.2145	0.0001	0.120260
CS4	0.6406	0.0855	56.1270	0.0001	0.124504
CS5	0.4156	0.0853	23.7433	0.0001	0.072711
CS7	0.1714	0.1100	2.4310	0.1190	0.042745
CS8	0.0152	0.1344	0.0129	0.9097	0.001832
TRAGECH1	0.3838	0.1394	7.5842	0.0059	0.040930
TRAGECH2	0.2116	0.0788	7.2157	0.0072	0.045767
TRAGECH4	-0.4384	0.0805	29.6728	0.0001	-0.088235
TRAGECH5	-0.6641	0.1004	43.7866	0.0001	-0.133999
TRAGECH6	-1.0963	0.1350	65.9539	0.0001	-0.197700
TRAGECH7	-1.8836	0.1550	147.6396	0.0001	-0.315675
STALOGN1	-0.0405	0.0721	0.3165	0.5737	-0.010496
STALOGN2	-0.2672	0.0846	9.9842	0.0016	-0.063960
STALOGN4	0.0531	0.1078	0.2426	0.6223	0.007385
C0	-0.2381	0.0710	11.2552	0.0008	-0.059216
C1	-0.2535	0.0760	11.1328	0.0008	-0.052424
C2	-0.1691	0.0820	4.2524	0.0392	-0.031389
C4	-0.0833	0.0835	0.9958	0.3183	-0.015739
NBPERS2	-0.1626	0.0776	4.3897	0.0362	-0.041260
NBPERS3	-0.3021	0.0896	11.3635	0.0007	-0.065184
NBPERS4	-0.2795	0.0968	8.3294	0.0039	-0.059708
NBPERS5	-0.5912	0.1077	30.1152	0.0001	-0.102090
REVENSN1	-2.2615	0.2515	80.8469	0.0001	-0.519338
REVENSN2	-1.9390	0.2529	58.7829	0.0001	-0.355879
REVENSN3	-1.4948	0.2478	36.3967	0.0001	-0.329743
REVENSN4	-1.0052	0.2452	16.8078	0.0001	-0.251054
REVENSN5	-0.5007	0.2484	4.0624	0.0438	-0.096900
DETTEN2	0.5863	0.0645	82.6155	0.0001	0.144417
DETTEN3	0.8675	0.0901	92.8010	0.0001	0.207042

Association of Predicted Probabilities and Observed Responses

Concordant = 79.2%		Somers' D =	0.586
Discordant = 20.6%		Gamma =	0.587
Tied = 0.2%		Tau-a =	0.293
(20177469 pairs)		c =	0.793

Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	3537	1124	4661
	NO EVENT	1416	2913	4329
	Total	4953	4037	8990

Sensitivity= 75.9% Specificity= 67.3% Correct= 71.7%
False Positive Rate= 28.6% False Negative Rate= 27.8%

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

ANNEXE 3

En analyse discriminante :

CS (Catégorie socioprofessionnelle de la personne de référence) :

Modalités :

3 : Cadres et professions intellectuelles supérieures
4 : Professions intermédiaires
5 : Employés
6 : Ouvriers
2 : Artisans, commerçants et chefs d'entreprise
8 : Autres personnes sans activité professionnelle
1 : Agriculteurs exploitants
7 : Retraités

TRAGECH (Tranche d'âge de la personne de référence) :

Modalités :

2 : De 25 à moins de 35 ans
1 : Moins de 25 ans
4 : De 45 à moins de 55 ans
3 : De 35 à moins de 45 ans
5 : De 55 à moins de 65 ans
6 : De 65 à moins de 75 ans
7 : 75 ans et plus

NBPERS (Nombre de personnes du ménage) :

Modalités :

4 : 4 personnes
3 : 3 personnes
2 : 2 personnes
5 : 5 personnes ou plus
1 : 1 personne

STALOGN (Statut d'occupation du logement) :

Modalités :

2 : Accédant à la propriété
4 : Logé gratuitement
3 : Locataire, sous-locataire
1 : Propriétaire ou copropriétaire

C (Catégorie de commune) :

Modalités :

4 : Complexe de l'agglomération parisienne (yc Paris)
3 : Unités urbaines de 100 000 habitants et plus
0 : Communes rurales
1 : Unités urbaines de moins de 20 000 habitants
2 : Unités urbaines de 20 000 à moins de 100 000 habitants

REVENSN (Revenu mensuel par tranche de tous les ménages) :

Modalités :

5 : De 18 000 F à moins de 37 000 F
6 : 37 000 F et plus
4 : De 11 000 F à moins de 18 000 F
3 : De 7 800 F à moins de 11 000 F
2 : De 5 800 F à moins de 7 800 F
1 : Moins de 5 800F

DETTEN (Montant total des crédits payés par le ménage au cours des 12 derniers mois) :

Modalités :

3 : 25 000 F et plus
2 : Ayant une dette de moins de 25 000 F
1 : Pas de dette

ANNEXE 3 (suite et fin)

En logistique

CS (Catégorie socioprofessionnelle de la personne de référence) :

Modalités :

3 : Cadres et professions intellectuelles supérieures
4 : Professions intermédiaires
5 : Employés
1 : Agriculteurs exploitants
7 : Retraités
2 : Artisans, commerçants et chefs d'entreprise
8 : Autres personnes sans activité professionnelle
6 : Ouvriers

TRAGECH (Tranche d'âge de la personne de référence) :

Modalités :

1 : Moins de 25 ans
2 : De 25 à moins de 35 ans
3 : De 35 à moins de 45 ans
4 : De 45 à moins de 55 ans
3 : De 35 à moins de 45 ans
5 : De 55 à moins de 65 ans
6 : De 65 à moins de 75 ans
7 : 75 ans et plus

NBPERS (Nombre de personnes du ménage) :

Modalités :

1 : 1 personne
2 : 2 personnes
4 : 4 personnes
3 : 3 personnes
5 : 5 personnes ou plus

STALOGN (Statut d'occupation du logement) :

Modalités :

4 : Logé gratuitement
3 : Locataire, sous-locataire
1 : Propriétaire ou copropriétaire
2 : Accédant à la propriété

C (Catégorie de commune) :

Modalités :

3 : Unités urbaines de 100 000 habitants et plus
4 : Complexe de l'agglomération parisienne (yc Paris)
0 : Communes rurales
2 : Unités urbaines de 20 000 à moins de 100 000 habitants
1 : Unités urbaines de moins de 20 000 habitants

REVENSN (Revenu mensuel par tranche de tous les ménages) :

Modalités :

6 : 37 000 F et plus
5 : De 18 000 F à moins de 37 000 F
4 : De 11 000 F à moins de 18 000 F
3 : De 7 800 F à moins de 11 000 F
2 : De 5 800 F à moins de 7 800 F
1 : Moins de 5 800F

DETTEN (Montant total des crédits payés par le ménage au cours des 12 derniers mois) :

Modalités :

3 : 25 000 F et plus
2 : Ayant une dette de moins de 25 000 F
1 : Pas de dette