Méthodologie statistique

M 2016/01

Le modèle Logit

Théorie et application

Cédric Afsa

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2016/01

Le modèle Logit

Théorie et application

Cédric Afsa *

Ce document a bénéficié des commentaires, corrections et remarques de Pauline Givord, Marine Guillerm et Olivier Sautory, que je remercie tout particulièrement.

Je reste responsable des erreurs qui subsisteraient.

* DEPP (Département de l'Évaluation, de la Prospective et de la Performance)
Ministère de l'Éducation Nationale, de l'Enseignement Supérieur et de la Recherche.

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101
18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 33 - Fax : 33 (1) 41 17 66 33 - CEDEX - E-mail :-DG75-L001@insee,fr - Site Web Insee : http://www.insee.fr

Le modèle Logit : Théorie et applications

Cédric Afsa *

Résumé

Le modèle logit a une double nature. D'une part, c'est un modèle de régression où la variable dépendante est binaire. D'autre part, c'est une méthode alternative à l'analyse discriminante linéaire. Par ailleurs, le modèle logit peut aussi être considéré comme un modèle économique de choix discrets.
L'objectif de ce document est double. D'abord, il passe en revue les caractéristiques du modèle et à cette occasion rappelle certaines notions de base comme la méthode d'estimation ou les tests d'hypothèse. Ensuite, il est appliqué à des données sur l'éducation, et un point particulier est fait sur la manière de présenter les résultats.

Mots clés : Modèle Logit ; régression logistique ; variable dichotomique

Abstract

The logit model has a dual nature. On the one hand it refers to a regression model where the dependent variable is binary. On the other hand it is an alternative to linear discriminant analysis. Moreover logit model may be considered as a discrete choice economic model.
The aim of the document is two-fold. Firstly key features of the logit model are presented and on this occasion basic notions such as estimation method or hypothesis testing are recalled. Secondly the model is applied to data on education and in particular stresses on how to present results.

KeyWords : Logit model ; logistic regression ; dichotomous variable

* DEPP (Département de l'Evaluation, de la Prospective et de la Performance)
cedric.afsa@education.gouv.fr

Sommaire

Avant-propos

Supposons que l'on sache distinguer, au sein d'une population, deux catégories d'individus. Par exemple, il y a sur le marché du travail les personnes en emploi et celles qui en recherchent un. Autre exemple : une partie des élèves étudie dans des établissements publics, l'autre est scolarisée dans le privé. Ou encore : parmi les candidats à un examen, les uns échouent, les autres réussissent. On part du principe, (quasiment) toujours vérifié, que les individus des deux catégories ne se ressemblent pas. On aimerait alors répondre à deux questions : sur quelles caractéristiques se différencient-ils ? et lesquelles jouent les premiers rôles en la matière ?

Le modèle logit [1] est tout à fait adapté à cette problématique. Outre qu'il permet d'identifier les caractéristiques distinguant les individus des deux groupes, il mesure aussi l'influence de chacune d'entre elles dans cette distinction.

Pour illustrer le propos, intéressons-nous à la question de l'accès à l'emploi sur le marché du travail. On cherche à connaître les facteurs qui font que certains individus ont plus de difficultés que d'autres à trouver un emploi. On distingue donc ceux qui sont en emploi et ceux qui en recherchent un. On souhaite plus précisément étudier le rôle joué en la matière par le critère de nationalité. On sait que les travailleurs étrangers ont davantage de problèmes d'emploi que leurs homologues français. Mais on sait aussi que, d'une manière générale, ces travailleurs ont un niveau de formation moins élevé que les français, donc une moindre qualification, ce qui les handicape sur le marché du travail. On peut dès lors se demander si les problèmes d'insertion dans l'emploi qu'ils rencontrent ne sont pas dus au moins en partie à la différence de qualification. S'il n'y a pas là ce qu'on appelle un *effet de structure* : le fait que les étrangers s'insèrent plus difficilement peut s'expliquer en partie par la différence *structurelle* des deux sous-populations en niveaux de qualification. On parle aussi d'*effet de composition*.

Pour le savoir, on peut conduire l'exercice consistant à se placer dans la situation – fictive – où les étrangers seraient autant formés que les Français. La nationalité aurait-elle encore un rôle dans l'accès à l'emploi ? Si oui, reste-t-il important ou non ? Le modèle logit permet précisément de faire l'exercice, en tenant compte à la fois de la nationalité et du niveau de formation, mesuré par exemple par le diplôme. On peut approfondir l'analyse et introduire d'autres caractéristiques comme le sexe, l'âge, le lieu de résidence, ..., c'est-à-dire créer une situation fictive où les Français

1. Sans autre précision, il s'agit du modèle logit *dichotomique*, qui modélise l'appartenance à une catégorie parmi *deux* possibles. A partir de trois catégories possibles (par exemple inactif/chômeur/en emploi), on parle de modèle logit *polytomique*.

et les étrangers auraient aussi la même pyramide des âges, la même structure par sexe, seraient répartis pareillement sur le territoire national, ..., puis regarder si les situations vis-à-vis de l'emploi seraient encore différentes.

Bien que le modèle logit soit aujourd'hui largement utilisé, il reste paradoxalement assez méconnu. On ignore souvent qu'il peut servir plusieurs finalités. Selon les situations rencontrées, on l'utilisera comme outil à visée ouvertement descriptive (analyse discriminante), ou bien comme modèle explicatif, sans parler de son apport à la modélisation économique des comportements individuels. Dans ce contexte et dans certains cas, il faut rester très vigilant [2] lorsqu'on interprète et commente ses résultats. Par exemple, il arrive trop souvent que des expressions telles que « toutes choses égales par ailleurs » ou « effet propre » (d'une caractéristique) soient utilisées à mauvais escient. C'est un point sur lequel on insistera à plusieurs reprises.

Le document se partage en deux grandes parties. La première présente, avec un formalisme minimal mais nécessaire, le logit dichotomique : sa spécification, la méthode d'estimation de ses paramètres, les indicateurs de qualité, les tests sur les paramètres, l'évaluation de l'importance de chaque facteur. On en profitera pour rappeler, dans des termes les plus simples possibles, certaines notions fondamentales (la définition d'une probabilité conditionnelle et le sens qu'on doit lui attribuer, le principe de l'estimation par le maximum de vraisemblance, la démarche à suivre pour tester une hypothèse, ...).

La seconde partie, plus pratique, est consacrée au traitement d'un exemple. Les principales étapes sont passées en revue, notamment la sélection et la préparation des variables, l'estimation des paramètres, la présentation des résultats. Les programmes SAS – procédures et macros écrites spécifiquement – sont présentés *in extenso*. Dans la mesure du possible, on fournira un certain nombre de conseils afin que les résultats puissent être compris par un lecteur ne connaissant pas *a priori* cet outil d'analyse.

Il n'est pas nécessaire de lire intégralement la première partie du document avant de passer à l'exemple d'application. On peut en faire une lecture sélective, puis se reporter aux pages 49 et suivantes, quitte à revenir, grâce aux renvois régulièrement faits, à la partie I du document pour approfondir certains points.

2. Le panneau signale des aspects délicats du modèle et de son utilisation, qu'il convient de traiter avec soin.

I. Le modèle Logit : un peu de théorie

I.1 La spécification du modèle : les différentes approches

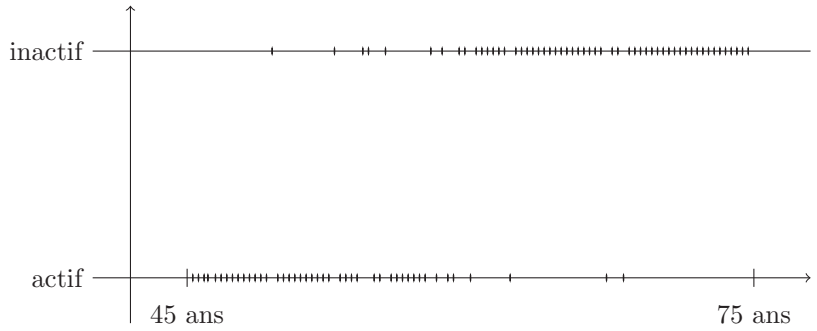
I.1.a Approche « descriptive »

On observe un échantillon d'individus dont on connaît K de leurs caractéristiques, représentées par les K variables $x_1, x_2, ..., x_K$. .

On suppose que les individus sont répartis en 2 catégories C_0 et C_1. Sur le marché du travail par exemple, certains travaillent (font partie de la catégorie C_1 des personnes en emploi), d'autres pas (catégorie C_0 des personnes sans emploi). Autre exemple, une partie des élèves de terminale a réussi les épreuves du baccalauréat (ils appartiennent à la catégorie C_1 des bacheliers), l'autre a échoué (catégorie C_0 des non bacheliers).

On souhaite analyser et quantifier le lien existant entre les caractéristiques individuelles x_k et l'appartenance à C_0 ou C_1. Il faut un outil – un modèle – spécifique pour pouvoir le faire. L'exemple suivant – très simplifié – va le montrer.

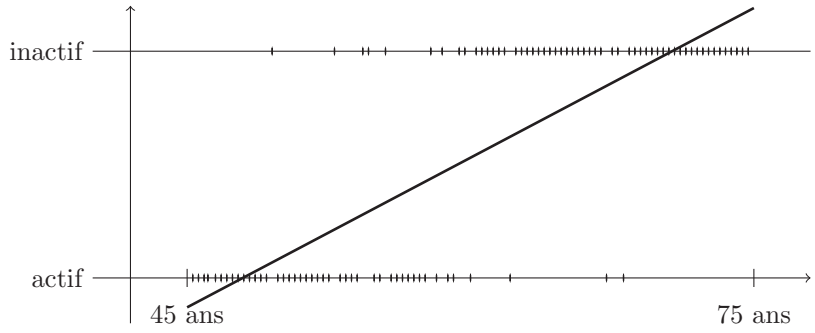
On a conduit une enquête auprès d'une centaine de personnes pour analyser le lien entre l'âge et l'activité sur le marché du travail. On s'intéresse plus précisément aux individus âgés de 45 ans à 75 ans. On leur a demandé de préciser s'ils étaient actifs ou inactifs. La figure ci-dessous représente les réponses individuelles à l'enquête.



Chaque point figure un individu. S'il a répondu être actif, il se situe sur la droite horizontale `actif`. Dans le cas contraire, il est sur la droite `inactif`. Les individus proches de 45 ans sont tous actifs, ceux proches de 75 ans sont tous inactifs. Il y a un lien positif entre l'âge et l'inactivité : le nombre de points sur la droite `inactif` (resp. `actif`) augmente (resp. diminue) avec l'âge. On s'en doutait. Plus intéressante est la question de savoir si ce lien est faible, moyen, fort, ..., en deux mots la question de sa quantification : de combien augmente l'inactivité quand on vieillit d'un an ?

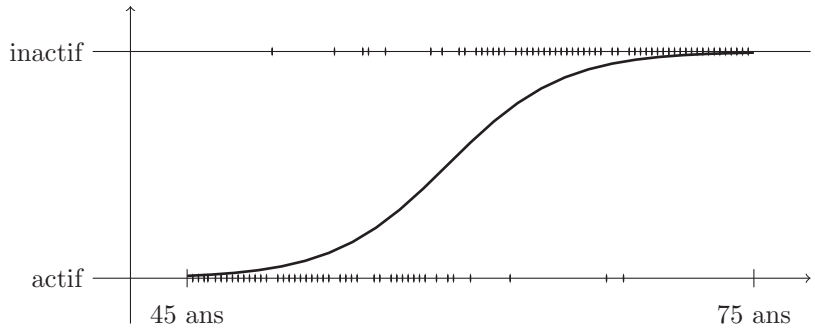
Poser ainsi la question suggère l'utilisation d'un outil comme la régression linéaire : on « explique » l'inactivité par l'âge, et la valeur estimée du paramètre associé à l'âge donne la force du lien. On procède donc comme suit. On crée la variable « à expliquer », nommée par exemple *inactif*, qui vaut 1 si la personne a répondu être inactive au moment de l'enquête (son inactivité est de 100%), et vaut 0 si elle se

dit active (son inactivité est de 0%) [3]. La figure suivante représente la droite de régression, celle qui passe « le plus près possible de tous les points ».



Cette manière de faire soulève au moins deux problèmes. Le premier est qu'on ne sait pas ce que représente chaque point de la droite, étant donné que la variable « à expliquer » prend deux valeurs et deux seulement. De plus, la valeur prédite de l'inactivité (qui se situe sur la droite de régression) est négative pour des âges proches de 45 ans. Il faut donc trouver une autre méthode.

Au lieu de s'intéresser au statut binaire inactif/actif, on se centre sur la *probabilité* d'être inactif. Il s'agit là d'une variable susceptible de varier continûment entre 0 et 1. On modélise alors le lien entre la probabilité d'être inactif et l'âge, et non entre le statut et l'âge. Puisque cette probabilité doit être comprise entre 0 et 1, son lien avec l'âge ne peut être représenté par une droite, mais par une courbe respectant cette contrainte. La figure suivante en est un exemple.



Cela posé, il faudrait définir précisément la relation fonctionnelle entre l'âge et l'inactivité de manière à pouvoir calculer la probabilité d'être inactif pour chaque

3. Ce type de modèle, appelé *modèle linéaire de probabilité*, est parfois utilisé lorsqu'il est légitime de le faire.

âge compris entre 45 et 75 ans. Pour ce faire, on a besoin d'un cadre formel général, exposé ci-dessous.

On part donc du principe que la population que l'on étudie est scindée en deux catégories, C_0 et C_1 (dans l'exemple précédent, C_0 contient les actifs et C_1 les inactifs). On dispose d'un échantillon de n individus indicés par i, représentatifs de cette population. On connaît K caractéristiques de ces individus, mesurées par les variables x_1, x_2, …, x_K. Pour l'individu i, les K variables prennent les valeurs x_{1i}, x_{2i}, …, x_{Ki}.

On pose que la probabilité P que l'individu i (compte tenu de ses caractéristiques $x_{1i}, x_{2i}, ..., x_{Ki}$) appartienne à C_1 ou à C_0 est une fonction des $x_{1i}, x_{2i}, ..., x_{Ki}$. On précise un peu la relation fonctionnelle en supposant que les probabilités d'appartenance dépendent d'une *combinaison linéaire* des caractéristiques. Formellement, cela s'écrit :

$$\begin{cases} P(i \in C_0 | x_{1i}, \ldots, x_{Ki}) = G(\beta_0^0 + \beta_1^0 x_{1i} + \cdots + \beta_K^0 x_{Ki}) \\ P(i \in C_1 | x_{1i}, \ldots, x_{Ki}) = G(\beta_0^1 + \beta_1^1 x_{1i} + \cdots + \beta_K^1 x_{Ki}) \end{cases} \tag{1}$$

où G est une fonction qui sera définie ultérieurement et où les β_0^0, β_1^0, …, β_K^0 et les β_0^1, β_1^1, …, β_K^1 sont les coefficients des combinaisons linéaires. Ce sont les *paramètres* du modèle. On notera l'ajout des deux paramètres β_0^0 et β_0^1, qui sont appelés parfois paramètres du « terme constant ». Ils sont associés à la variable x_0 valant systématiquement 1. A ce stade, on a donc deux séries de paramètres β_k^j :

- la série β_0^0, β_1^0, …, β_K^0 associée à la catégorie C_0 $(j = 0)$;
- la série β_0^1, β_1^1, …, β_K^1 associée à la catégorie C_1 $(j = 1)$.

On verra plus loin que ces deux séries peuvent se « condenser » en une seule.

Avant de poursuivre, une remarque sur les notations. La combinaison linéaire des caractéristiques peut s'écrire de manière synthétique, pour $j = 0$ ou $j = 1$:

$$\beta_0^j + \beta_1^j x_{1i} + \cdots + \beta_K^j x_{Ki} = \begin{pmatrix} 1 & x_{1i} & \ldots & x_{Ki} \end{pmatrix} \begin{pmatrix} \beta_0^j \\ \beta_1^j \\ \vdots \\ \beta_K^j \end{pmatrix} = x_i \beta^j, \tag{2}$$

où $x_i = \begin{pmatrix} 1 & x_{1i} & \ldots & x_{Ki} \end{pmatrix}$ est le vecteur-*ligne* des caractéristiques de l'individu i et β^j le vecteur-*colonne*[4] des paramètres du modèle. On peut alors réécrire (1) de manière condensée :

$$P(i \in C_j | x_i) = G(x_i \beta^j) \quad \text{pour } j = 0, 1.$$

4. Il est préférable de représenter le vecteur des caractéristiques individuelles par un vecteur-ligne et celui des paramètres par un vecteur-colonne. On en verra l'avantage lors de l'écriture des programmes SAS.

Quelle fonction choisir pour G ? $P(i \in C_0|x_i)$ et $P(i \in C_1|x_i)$ étant des probabilités, on doit avoir :

$$\begin{cases} 0 < P(i \in C_0|x_i) < 1 \quad \text{et} \quad 0 < P(i \in C_1|x_i) < 1 \\ P(i \in C_0|x_i) + P(i \in C_1|x_i) = 1 \end{cases} \tag{3}$$

Poser $G(x_i\beta^j) = e^{x_i\beta^j}$ assurerait $P(i \in C_j|x_i) > 0$. Mais les autres contraintes ne seraient pas vérifiées. Pour qu'elles le soient, il suffit de normer les deux quantités $e^{x_i\beta^0}$ et $e^{x_i\beta^1}$, c'est-à-dire les diviser par leur somme. On obtient alors :

$$P(i \in C_0|x_i) = \frac{e^{x_i\beta^0}}{e^{x_i\beta^0} + e^{x_i\beta^1}} \qquad \text{et} \qquad P(i \in C_1|x_i) = \frac{e^{x_i\beta^1}}{e^{x_i\beta^0} + e^{x_i\beta^1}}$$

C'est cette forme fonctionnelle qui donne au modèle son nom de *logit*.

On peut simplifier en remarquant qu'une seule probabilité suffit pour le représenter, puisque la somme de $P(i \in C_0|x_i)$ et de $P(i \in C_1|x_i)$ est égale à 1. L'une se déduit de l'autre. On se centre sur la probabilité d'appartenir à C_1. Elle s'écrit :

$$P(i \in C_1|x_i) = \frac{e^{x_i\beta^1}}{e^{x_i\beta^0} + e^{x_i\beta^1}} = \frac{1}{1 + e^{x_i(\beta^0 - \beta^1)}}$$

Finalement, si on pose $\beta = \beta^1 - \beta^0$, on a :

$$P(i \in C_1|x_i) = \frac{1}{1 + e^{-x_i\beta}} \tag{4}$$

Dans le cas d'une seule variable x_1, on peut représenter la courbe, donnée par l'équation (4), sur un plan, avec en ordonnée la probabilité d'appartenir à la catégorie C_1 et en abscisse les valeurs prises par la variable x_1. C'est ce qui a été fait page 8, où la catégorie C_1 est celle des inactifs et la variable x_1 est l'âge de la personne enquêtée [5].

L'équation du modèle s'écrit plus fréquemment avec la variable catégorielle y définie par : $y_i = 1$ si $i \in C_1$ et $y_i = 0$ si $i \in C_0$. La formulation (4) devient :

$$\boxed{P(y_i = 1|x_i) = \frac{1}{1 + e^{-x_i\beta}}} \tag{5}$$

C'est elle qui est très généralement utilisée. Dans cette expression, les valeurs prises par les variables y_i et x_i sont connues puisqu'observées sur l'échantillon d'étude. En revanche, les valeurs des paramètres $(\beta_0, \ldots, \beta_K) = \beta$ sont inconnues. On verra par la suite (pages 25 et suivantes) comment les obtenir.

Une remarque sur les hypothèses du modèle. Celle imposant que la probabilité d'appartenance soit fonction d'une combinaison linéaire des caractéristiques – hypothèse dite d'*additivité* – n'est pas innocente. C'est elle qui permet d'évaluer le

5. Très précisément, la courbe a été dessinée avec les valeurs $\beta_0 = -19.4$ et $\beta_1 = 0.33$.

rôle de chaque variable x_k dans l'appartenance à l'une ou l'autre catégorie, *indépendamment des autres variables*. Pour voir ce que cela signifie, reprenons l'exemple du marché du travail, où sont distinguées les personnes en emploi ($j = 1$) et celles sans emploi ($j = 0$). Les caractéristiques individuelles sont le sexe x_1, le niveau de formation x_2, l'âge x_3 et la nationalité x_4. La variable x_4 vaut 0 ou 1 selon que l'individu est de nationalité française ou étrangère. Fixons les trois autres variables à des valeurs quelconques, par exemple celles les plus fréquemment rencontrées dans l'échantillon. Si on connaît les valeurs des paramètres, on peut alors calculer, grâce à la formule (5), les deux probabilités d'appartenance à la catégorie C_1 correspondant aux deux valeurs possibles de x_4. La différence entre ces deux probabilités mesure le rôle joué par le critère de nationalité dans l'appartenance à C_1, à âge, sexe et niveau de formation fixés ou constants.

Ainsi, l'hypothèse d'additivité permet d'évaluer l'impact, sur la probabilité d'appartenir à C_1, de la variation de chaque variable x_k, les autres étant maintenues constantes.

Autre remarque : il faut écrire $P(y_i = 1|x_i)$ et non simplement $P(y_i = 1)$. L'écriture adoptée rappelle que la quantité $P(y_i = 1|x_i)$ dépend bien de x, comme le montre le membre de droite de l'expression (5). La quantité est une *probabilité conditionnelle*, au sens où elle mesure la probabilité que y_i soit égal à 1 conditionnellement aux (i.e. compte tenu des) variables x_1, x_2, ..., x_K introduites dans le modèle. Si on ajoute une variable x_{K+1} à la liste, alors la probabilité change. Il s'agit là d'un point très important, sur lequel on aura l'occasion de revenir.

I.1.b Une application particulière : le contraste logistique

La courbe, comme celle de la page 8, dérivée de l'équation (5) avec une seule variable x_1 est donc bien adaptée à la représentation d'une probabilité et de sa variation selon différentes valeurs de la variable x_1. Cette forme fonctionnelle permet de résoudre le problème de la comparabilité d'évolutions temporelles de pourcentages, problème qui se pose dans les termes suivants.

Supposons que l'on suive, sur longue période et dans une population de taille constante, la diffusion d'un produit nouveau en la mesurant par l'évolution du taux d'équipement de la population en ce produit. Passer de 5% à 10% correspond à une augmentation de 5 points du taux. On a toutefois le sentiment que cette évolution est plus importante que celle qui fait passer de 50% à 55% où l'écart est également de 5 points. En effet, dans le premier cas le nombre de personnes équipées est multiplié par 2, alors que dans le second cas l'augmentation relative est de 10%. Les deux progressions sont donc jugées équivalentes si on raisonne en écart absolu (c'est-à-dire avec une échelle additive), mais si on raisonne en écart relatif (c'est-à-dire avec une échelle mutiplicative), la progression de 5% à 10% est jugée beaucoup plus importante. Que conclure ? L'échelle logistique, qui est adaptée à cette question de diffusion d'une innovation [6], permet de trancher.

6. L'origine du modèle *logit* remonte au XIXème siècle, lorsque Pierre-François Verhulst publia

Soit P la proportion des personnes possédant un bien donné (ou toute autre caractéristique, comme un diplôme). Cette proportion évolue avec le temps t : $P = P(t)$. Dans le cas d'un nouveau bien, elle est nulle juste avant sa mise sur le marché, puis augmente, d'abord faiblement, dès que le bien est disponible. Soit y la variable indiquant si l'individu dispose du bien ($y = 1$) ou non ($y = 0$). La probabilité, pour une personne quelconque, de posséder le bien à l'instant t n'est rien d'autre que la proportion $P(t)$: $P(y = 1|t) = P(t)$.

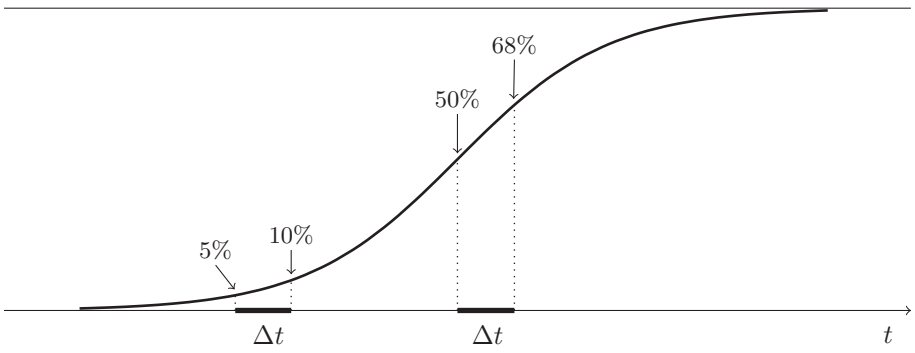
On fait dépendre la proportion P du temps t selon la relation :

$$P = P(t) = \frac{1}{1 + e^{-t}} \tag{6}$$

On retrouve l'équation (5) du modèle *logit* avec une seule variable x, qui est ici le temps t. La relation (6) permet d'exprimer t en fonction de P. On a en effet, tous calculs faits :

$$t = \ln \frac{P}{1 - P} \tag{7}$$

où ln est le logarithme népérien.



La figure ci-dessus représente l'évolution temporelle de P selon la relation (6). Après la mise sur le marché du nouveau bien, la proportion croît d'abord très modérément, puis plus rapidement ensuite, pour de nouveau augmenter lentement au moment où le marché arrive à saturation (i.e. la grande majorité des personnes sont équipées). Il faut une durée Δt pour que le taux d'équipement passe de 5% à 10%. Lorsque la moitié des personnes possède le bien ($P = 50\%$), la diffusion à d'autres est plus rapide si bien que sur la même durée Δt, la proportion augmente davantage, de 18 points pour être précis. De ce point de vue, l'augmentation de 50% à 68% est équivalente à l'augmentation de 5% à 10%.

D'une manière générale, soit P_1 (resp. P_2) la valeur de P atteinte à l'instant t_1 (resp. t_2). L'intervalle de temps nécessaire pour que la proportion P passe de P_1 à

en 1838 un article qui présente la fonction logistique comme outil de description de la croissance de populations (voir J.S. Cramer (2002), « The origins and development of the logit model », *Tinbergen Institute Discussion Paper*, n° 199/4).

P_2 est, en vertu de la relation (7), égale à :

$$t_2 - t_1 = \ln \frac{P_2}{1 - P_2} - \ln \frac{P_1}{1 - P_1}$$

Cette différence est appelée *contraste logistique*

Un exemple d'application. Le taux de bacheliers dont les parents sont cadres ou exercent une profession intermédiaire est passé de 63% dans la génération 64-68 à 84% dans la génération 84-88 [7]. Sur la même période, le taux concernant les enfants d'ouvriers ou d'employés est passé de 22% à 55%.

Le contraste logistique vaut $1,13$ [8] pour les enfants de cadres. Pour les enfants d'ouvriers, il est plus élevé, car il vaut $1,47$ [9]. Il faut ainsi plus de temps pour passer de 22% à 55% que pour passer de 63% à 84%. Or l'évolution du taux de bacheliers pour les enfants de cadres et pour les enfants d'ouvriers a eu lieu sur le même laps de temps, c'est-à-dire sur les 20 années qui séparent les générations 64-68 et 84-88. En conséquence, la diffusion du baccalauréat chez les enfants de familles ouvrières s'est faite à un rythme accéléré, comparativement à celle des enfants de cadres. Les inégalités sociales devant le baccalauréat ont donc diminué sur la période considérée.

I.1.c Approche « explicative »

L'approche présentée en section I.1.a et qu'on a qualifiée de « descriptive », est avant tout pragmatique. La forme fonctionnelle de G est définie de manière *ad hoc*, pour respecter les propriétés (3) des probabilités $P(i \in C_0|x_i)$ et $P(i \in C_1|x_i)$. La seconde approche, dite ici « explicative », a une nature un peu plus théorique.

Pour l'introduire, on prend l'exemple de la réussite à un examen (baccalauréat ou autre). A l'issue des épreuves, on peut distinguer deux catégories d'élèves : les admis et les recalés. Cette manière de présenter les choses – parler de catégories d'élèves – relève de l'approche « descriptive », vue précédemment. L'approche plus « explicative » de la question consiste à la traiter de la manière suivante.

La réussite ou l'échec à l'examen sont supposés révéler le niveau de l'élève. Notons-le y^*. On ne le connaît pas. On sait seulement que le candidat a réussi ou a échoué. Notons y la variable binaire indiquant l'issue de l'examen : elle vaut 1 en cas de réussite et 0 en cas d'échec. On établit le lien suivant entre y^* et y : dire qu'un élève passe son examen avec succès, c'est dire que son niveau est supérieur à un certain seuil s_0. Par conséquent, le lien entre les deux variables se formalise ainsi :

$$\begin{cases} y_i = 1 \ \Leftrightarrow \ y_i^* > s_0 \\ y_i = 0 \ \Leftrightarrow \ y_i^* < s_0 \end{cases}$$

On dispose par ailleurs de plusieurs informations sur les caractéristiques socio-

7. Voir *L'état de l'école*, DEPP, édition 2012, page 69
8. C'est-à-dire : $\ln(84/16) - \ln(63/37) = 1,658 - 0.532 = 1,126$
9. C'est-à-dire : $\ln(55/45) - \ln(22/78) = 0,201 - (-1,266) = 1,467$

démographiques des élèves : on connaît les valeurs prises par un ensemble de variables $x_1, x_2, ..., x_K$ pour chaque élève.

La question centrale est ici de savoir si les caractéristiques de ces élèves influent sur leur niveau, si elles sont susceptibles de l'expliquer et dans quelle mesure. Formalisons tout cela avec le modèle le plus simple traduisant l'influence des variables x_k (pour $k = 1, \ldots, K$) sur y^*, celui où les effets des variables explicatives $x_1, x_2, ..., x_K$ sur le niveau de compétences sont supposés être additifs. Ce modèle s'écrit :

$$y_i^* = \beta_0' + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \ldots + \beta_K x_{Ki} + u_i \tag{8}$$

Dans l'équation (8), le paramètre β_k représente l'effet de la variable x_k sur le niveau y^*. On remarquera l'ajout du terme résiduel u. Il contient notamment toutes les informations, toutes les variables qui peuvent influer sur le niveau de l'élève mais qui nous restent inconnues car nous ne les observons pas. Certaines sont dites inobservées lorsqu'elles ne figurent pas dans la source de données mais pourraient y être, d'autres restent inobservables, car en pratique elles le sont (par exemple, l'état de stress de l'élève le jour de l'examen).

L'équation (8) ressemble à une régression linéaire, à ceci près que la variable dépendante y^* est latente, c'est-à-dire qu'elle n'est pas observée. Ceci justifie l'appellation « modèle à variable latente » parfois donnée à (8). Puisqu'on ne connaît pas y^*, on ne peut pas estimer β comme on pourrait le faire avec un modèle de régression linéaire. Il faut donc aller plus loin dans la spécification du modèle.

On pose alors deux hypothèses supplémentaires. La première consiste à considérer u comme variable aléatoire et à supposer que sa distribution est symétrique. On note G sa fonction de répartition. La seconde hypothèse, plus contraignante comme on le verra, est l'indépendance de la variable u et des variables x_k. Cela s'écrit formellement :

$$u \perp\!\!\!\perp x_k \quad \forall\, k = 1, \ldots, K. \tag{9}$$

Cette hypothèse implique que la *probabilité conditionnelle de u sachant x*, notée $P(u|x)$, est indépendante de x.

Pour bien comprendre cette propriété d'indépendance, on peut se représenter la probabilité conditionnelle de u sachant x de la manière suivante. On fixe les valeurs des K variables composant le vecteur x et on se restreint au sous-ensemble défini par les individus dont les K caractéristiques ont les valeurs qu'on vient de fixer. Soit $\mathcal{E}(x)$ ce sous-ensemble. La probabilité de u conditionnelle à cette valeur particulière du vecteur x représente alors la manière dont varie u dans le sous-ensemble $\mathcal{E}(x)$. Prenons maintenant différentes valeurs de x. Il leur correspond autant de sous-ensembles $\mathcal{E}(x)$. Il n'y a pas de raison de penser que u varie de la même manière dans les différents sous-ensembles $\mathcal{E}(x)$. Autrement dit, la probabilité de u conditionnelle à x, c'est-à-dire la manière dont u varie dans $\mathcal{E}(x)$, dépend de x. Sauf dans le cas où u et x sont indépendants : la variable u varie alors de la même manière dans *tous*

les sous-ensembles $\mathcal{E}(x)$. En d'autres termes, la probabilité $P(u|x)$ ne dépend pas de x, elle est égale à $P(u)$. On va utiliser ce résultat un peu plus tard.

On poursuit donc la spécification du modèle. A défaut d'observer y^*, on se reporte sur la variable y qui, elle, est observée. On s'intéresse alors aux fréquences des réussites ($y = 1$) et des échecs ($y = 0$) pour différentes valeurs de x et on regarde si ces fréquences varient sensiblement selon x. On est ainsi amené à examiner les probabilités $P(y = 1|x)$ et $P(y = 0|x)$. On a :

$$\begin{aligned} P(y = 1|x) = P(y^* > s_0|x) &= P(\beta'_0 - s_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u > 0|x) \\ &= P(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u > 0|x), \end{aligned}$$

avec $\beta_0 = \beta'_0 - s_0$ [10]. En utilisant la notation (2) de la section précédente, la dernière expression s'écrit de manière plus condensée : $P(x\beta + u > 0|x)$. Il vient alors :

$$P(y = 1|x) = P(x\beta + u > 0|x) = P(-u < x\beta|x) = P(-u < x\beta)$$

La dernière égalité provient de l'indépendance de u et de x, c'est-à-dire du fait que la probabilité $P(u|x)$ ne dépend pas de x, ce qui permet de supprimer le conditionnement par x, comme on l'a vu plus haut. On poursuit :

$$P(y = 1|x) = P(-u < x\beta) = P(u < x\beta)$$

puisque la loi de u est supposée être symétrique. Finalement :

$$P(y = 1|x) = G(x\beta) \tag{10}$$

où G est la fonction de répartition de la loi de u [11].

Il reste à définir la fonction G, c'est-à-dire à choisir la loi de probabilité de u. Il y a deux possibilités. La première est la loi *logistique*. Sa particularité est qu'il n'y a pas de représentation analytique directe de sa fonction de densité, c'est-à-dire qu'on ne peut pas écrire de formule représentant a priori la probabilité $P(u)$. En revanche, on sait écrire sa fonction de répartition. Elle est égale à $G(a) = \frac{1}{1+e^{-a}}$. Dans ces conditions, (10) devient :

$$P(y = 1|x) = \frac{1}{1 + e^{-x\beta}} \tag{11}$$

On retrouve l'expression (5) du modèle logit. Notons ici que le modèle est parfois appelé *régression logistique*. Cela provient du fait que (11) est dérivée du modèle de régression (8), qui est à proprement parler une régression linéaire à résidus logistiques.

10. Cette égalité traduit le fait qu'on ne peut pas *identifier* le seuil minimal s_0 et donc en estimer le niveau.

11. La valeur que prend au point a la fonction de répartition de la loi de u est, rappelons-le, la probabilité que u soit inférieure à a.

La seconde possibilité est de faire suivre à u la loi *normale centrée réduite*, dont la densité, traditionnellement notée $\phi(u)$, s'écrit analytiquement :

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{u^2}{2}]$$

La fonction de répartition est notée $\Phi(a)$, dont on ne connaît l'expression que sous la forme d'une intégrale. L'expression (10) se réécrit alors :

$$P(y = 1|x) = \Phi(x\beta)$$

Il s'agit du modèle *probit*.

I.1.d Comparaison des deux approches

Parler d'approche « descriptive » est d'autant plus justifié qu'il existe un lien étroit entre le modèle logit et l'analyse discriminante. Cette technique, rappelons-le, vise à décrire puis prédire, à partir des valeurs prises par plusieurs variables dites prédictives, l'appartenance d'un ensemble d'individus à des groupes prédéfinis. Par exemple, dans le domaine médical, on peut détecter les groupes à hauts risques cardiaques, c'est-à-dire prédire l'appartenance de patients à ces groupes à partir de leur poids, leur mode d'alimentation, leurs antécédents familiaux, leurs conduites à risque (consommation de tabac, d'alcool, ...).

Ce lien entre modèle logit et analyse discriminante a été formellement établi il y a plusieurs décennies [12]. Sous certaines conditions en particulier sur les variables x [13], on montre que l'analyse discriminante (linéaire) est un cas particulier du modèle logit.

La seconde approche, qualifiée d'« explicative », est très différente. Elle s'apparente aux analyses causales, qui cherchent à établir un lien de cause à effet entre la variable explicative principale, c'est-à-dire la variable x_k que l'on privilégie dans l'analyse, et la variable dite d'intérêt (ici, la variable binaire à expliquer). Le modèle de base pour cette approche est (8). L'effet de x_k sur y^* est mesuré par le paramètre β_k. Le problème posé par cette approche est que dans bien des cas il est difficile d'estimer correctement l'effet causal de la variable principale x_k. Cela se produit notamment lorsque la propriété (9) d'indépendance du résidu u et de x_k n'est pas satisfaite.

En guise d'illustration, reprenons l'exemple de la réussite à un examen comme variable révélatrice du niveau de l'élève. On cherche à l'expliquer par une relation

12. Parmi les premiers travaux en la matière, on citera G.W. Ladd, « Linear Probability Functions and Discriminant Functions », *Econometrica*, 1966, ou encore D. McFadden, « A Comment on Discriminant Analysis 'versus' Logit Analysis », *Annals of Economic and Social Measurement*, 1976.

13. Pour être précis, les valeurs effectivement prises par les variables x dans chacun des deux groupes doivent pouvoir être considérées comme des valeurs tirées dans des lois normales ayant la même matrice de variance-covariance. Voir à ce sujet O. Sautory et C. Vong, « Une étude comparative des méthodes de discrimination et de régression logistique », *Insee Méthodes*, n° 46-47-48, 1995.

du type (8). Supposons que la variable x_1 indique si l'élève étudie dans un établissement public ($x_1 = 0$) ou dans un établissement privé ($x_1 = 1$). On s'y intéresse particulièrement, car on aimerait connaître l'effet du secteur (public/privé) sur le niveau de l'élève, effet mesuré par le paramètre β_1. Si on l'estime par le modèle logit (11), on risque fort de se tromper et de récupérer une valeur du paramètre qui ne correspond pas au « vrai » effet causal de x_1 sur y^*. Car les élèves qui fréquentent les établissements privés ne sont pas comme tous les autres. Ils viennent plus souvent de familles aisées, pour qui le recours au privé fait partie d'une stratégie visant la meilleure réussite possible de l'enfant. Ces élèves ont pu aussi être sélectionnés sur leurs résultats scolaires avant d'entrer dans l'établissement. Ces critères distinctifs ont par ailleurs une influence déterminante dans les apprentissages et, par voie de conséquence, dans la réussite à l'examen. Les données les mesurant – l'attitude des parents, les compétences *ex ante* de l'élève – sont difficiles à collecter, si bien que ces informations, en règle trés générale, ne sont pas observées et font partie du résidu u. Dans ce cas, elles tirent à la hausse la valeur moyenne du résidu calculée sur la sous-population des élèves fréquentant les établissements privés (i.e. $E(u|x_1 = 1)$), par rapport à celle calculée sur les élèves du public (i.e. $E(u|x_1 = 0)$). On a donc : $E(u|x_1 = 1) > E(u|x_1 = 0)$. En conséquence, la probabilité conditionnelle de u sachant x_1 (voir la section I.1.a *supra*) dépend de x_1 puisque les valeurs moyennes de u pour $x_1 = 0$ et $x_1 = 1$ sont différentes. La propriété (9) n'est pas satisfaite. On dit dans ce cas que la variable x_1 est *endogène*. On ne peut donc pas passer de l'expression de base (8) du modèle à sa formulation logistique (11).

Que se passe-t-il si on estime β_1 comme paramètre d'un logit en l'absence d'informations telles que la stratégie parentale ou les résultats antérieurs de l'élève, maintenant implicitement l'hypothèse d'indépendance entre le résidu u de l'expression (8) et la variable x_1 ? Pour le voir, on repart du modèle :

$$y_i^* = \beta_0 + \beta_1 x_{1i} + x_{(K-1)\,i}\beta_{(K-1)} + u_i \tag{12}$$

où x_1 est la variable de secteur, $x_{(K-1)}$ les autres variables du modèle et $\beta_{(K-1)}$ leurs paramètres associés. On suppose donc qu'il manque dans $x_{(K-1)}$ les variables comme la stratégie parentale ou le niveau de l'élève. Estimer β_1 avec la spécification (12) suppose implicitement que u est indépendant de x_1. Ceci exclut que les variables manquantes soient comprises dans le résidu (sinon il y aurait un lien entre u et x_1). En conséquence, elles sont englobées dans la variable x_1 (puisqu'elles ne figurent pas dans $x_{(K-1)}$). Dans ces conditions, la valeur de β_1 estimée par (12) capte à la fois les effets – positifs – des deux variables manquantes sur y^*, et l'effet net du secteur privé ($x_1 = 1$) sur y^*. En d'autres termes, le paramètre β_1 ainsi estimé surestime l'effet « propre » du privé sur y^*, appelé aussi « effet causal ».

Il faut donc redoubler de prudence lorsqu'on commente les résultats d'un modèle *logit* « explicatif », ne pas parler d'« effet pur » d'une variable explicative lorsqu'elle est présumée être endogène. La formulation « toutes choses égales par ailleurs » n'est

guère plus satisfaisante. On y reviendra dans la seconde partie du document.

I.1.e Une troisième approche

Outre ces deux approches (« descriptive » et « explicative »), il faut en mentionner une troisième, même si elle ne sera pas traitée dans la suite du document. Elle est fondée sur la théorie économique standard des comportements individuels.

Supposons que l'individu i ait à choisir entre deux options possibles, notées 0 et 1. Sa décision s'appuie sur le modèle sous-jacent suivant, appelé *modèle d'utilité stochastique additive* (en anglais : *additive random utility model* – ARUM) : l'utilité U_{ji} qu'il retire (ou retirerait) de l'option j, où j peut prendre la valeur 0 ou 1, est la somme d'une composante « déterministe » V_{ji} et d'une composante aléatoire u_{ji} :

$$U_{ji} = V_{ji} + u_{ji}$$

La première est nommée ainsi car elle est entièrement déterminée ou expliquée par un ensemble de caractéristiques individuelles observées et notées x_i : $V_{ji} = V_j(x_i)$. La forme généralement retenue de la fonction $V_j(x)$ est linéaire en x : $V_j(x_i) = x_i\beta^j$, en utilisant la notation condensée (2) *supra*. La seconde composante rassemble les variables inobservées et inobservables qui peuvent jouer sur la décision de l'agent i. Elle est supposée varier de manière aléatoire. En résumé, l'utilité que i retire(rait) de l'option j s'écrit :

$$U_{ji} = x_i\beta^j + u_{ji} \tag{13}$$

La règle de décision est alors la suivante : l'individu choisit une des deux options si l'utilité qu'il en retire est supérieure à l'utilité attendue de l'autre option. Si y est la variable binaire repérant l'option choisie (i.e. $y_i = 0$ si l'individu i a choisi l'option 0, et $y_i = 1$ si i a choisi 1), alors :

$$\begin{cases} y_i = 0 \;\Leftrightarrow\; U_{0i} > U_{1i} \\ y_i = 1 \;\Leftrightarrow\; U_{1i} > U_{0i} \end{cases} \tag{14}$$

En introduisant les caractéristiques observées x et en remplaçant l'utilité par son expression (13), on a :

$$\begin{aligned} P(y_i = 1|x) &= P(V_1(x_i) + u_{1i} > V_0(x_i) + u_{0i}) \\ &= P(u_{0i} - u_{1i} < V_1(x_i) - V_0(x_i)) \\ &= P(u_{0i} - u_{1i} < x_i(\beta^1 - \beta^0)) \\ &= P(u_{0i} - u_{1i} < x_i\beta) \end{aligned}$$

où $\beta = \beta^1 - \beta^0$. Finalement :

$$P(y_i = 1|x) = G(x_i\beta) \tag{15}$$

où G est la fonction de répartition de la loi $u_0 - u_1$. On montre que si u_0 et u_1 suivent la loi dite « type I extreme-value » ou loi de Gumbel, dont la fonction de densité s'écrit $f(u) = e^{-u} \exp[-e^{-u}]$, alors on retrouve l'expression du modèle logit déjà rencontrée :

$$G(x_i\beta) = \frac{1}{1 + e^{-x_i\beta}}$$

Ce cadre théorique n'est pas toujours pertinent, loin s'en faut. Par exemple, cela n'a pas de sens de traiter ainsi la réussite à un examen, car le candidat ne choisit pas d'échouer ! *A priori*, l'utilité U_{1i} est toujours supérieure à U_{0i}. En revanche, il est surtout adapté aux cas où les deux options entre lesquelles l'individu doit trancher ont elles-mêmes des caractéristiques qui font partie des critères de décision. Le choix d'un mode de transport en est l'exemple-type. Supposons que l'individu i hésite entre deux moyens de transport pour se rendre dans une ville éloignée de son domicile : le train d'un côté, l'avion de l'autre. Pour arbitrer, il tiendra compte notamment des prix et des temps totaux du trajet, pour le train et pour l'avion.

Ce cas où des caractéristiques des options entrent dans les critères de décision se formalise de la manière suivante. Soit z_{ji}, pour $j = 0$ ou 1, ces caractéristiques pour l'individu i. Il s'agit par exemple de ce que lui coûtera(it) chacun des modes de transport, les temps de trajet respectifs qu'il connaîtra(it). En supposant qu'elles agissent de manière additive sur l'utilité, l'expression (13) devient :

$$U_{ji} = x_i\beta^j + z_{ji}\gamma + u_{ji} \tag{16}$$

En appliquant toujours la même règle de décision (14), la probabilité de prendre l'option 1 s'écrit maintenant :

$$P(y_i = 1|x_i, z_{0i}, z_{1i}) = G[x_i(\beta^1 - \beta^0) + (z_{1i} - z_{0i})\gamma] = \frac{1}{1 + e^{-x_i\beta - (z_{1i} - z_{0i})\gamma}}$$

Formalisé ainsi, le modèle, parfois appelé *modèle logit conditionnel* selon la dénomination que lui a donné McFadden (*conditional logit model*), est passablement différent du modèle *logit* représenté par les expressions (5) ou (11). Il contient en effet, en plus des caractéristiques individuelles x, des variables – les z_j – qui varient avec l'option proposée. De plus, et surtout, les z_j sont les variables du modèle à privilégier dans l'analyse. Les caractéristiques individuelles sont introduites d'abord pour prendre en compte l'hétérogénéité observée des individus. Car l'utilisation qui peut être faite de ce type de modèle est d'estimer l'impact sur les comportements d'une modification des tarifs. Par exemple, réduire de, mettons, 10 % en moyenne les prix des billets de train attirerait-il une partie de la clientèle prenant habituellement l'avion, et si oui dans quelle proportion ?

Une dernière remarque. Si les modèles d'utilité stochastique sont bien adaptés aux cas où les caractéristiques des options font partie des critères de choix, on peut à la

rigueur y faire référence lorsqu'elles ne sont pas mesurées. Supposons que l'on ait à modéliser un choix d'orientation, entre la voie générale et la voie professionnelle par exemple. Dans les critères de choix pourrait figurer ce que craint ou espère l'élève à l'issue de ses études (le taux de chômage qu'il risque de connaître, le salaire espéré), mais aussi le coût de sa scolarité qu'il s'attend à supporter. Si on dispose de ces informations, alors on peut spécifier puis estimer un modèle du type (16), où les variables z sont le taux de chômage, le salaire et le coût de scolarité attendus. En l'absence de données sur ces variables, on peut s'en tenir à (13), et on considère que les informations non disponibles font partie du résidu u [14]. Mais la portée du modèle reste limitée.

14. Cet exemple du choix d'orientation fait partie de ceux qu'on ne peut pas modéliser par une expression du type (8), car il n'existe pas de variable latente adaptée.

I.2 Les variables du modèle

Jusqu'à présent, nous nous sommes concentrés sur la formalisation du modèle pour qu'il soit adapté au caractère particulier de la variable catégorielle y. Nous n'avons donné aucune précision sur les autres variables du modèle, i.e. $x_1, x_2, \ldots, x_K$. Elles peuvent être de natures très différentes, dont il faut tenir compte pour les traiter de manière adéquate.

On distingue d'abord les variables dites numériques (ou quantitatives) et les variables qualitatives. Par exemple, l'âge de la personne ou le nombre d'habitants de sa commune de résidence sont des variables numériques. Le sexe, le diplôme ou la filière d'enseignement sont des variables qualitatives.

Ensuite, il y a plusieurs types de variables qualitatives. Premier type, les variables binaires (appelées aussi dichotomiques) qui, comme le sexe, ne comportent que deux modalités. Second type, les variables polytomiques, qui ont plus de deux modalités. On a coutume de distinguer parmi elles celles qui sont ordonnées et celles qui ne le sont pas. Par exemple, les diplômes, en règle générale, permettent de classer leurs détenteurs les uns par rapport aux autres. On dira ainsi que le baccalauréat est « supérieur » au brevet. En revanche, les filières d'enseignement ou encore les disciplines (sciences, lettre droit, ...) ne sont pas ordonnables.

I.2.a Les variables continues

Le cas le plus simple à traiter est celui des variables numériques, qui sont introduites telles quelles dans le modèle. Supposons par exemple qu'on analyse l'influence de la seule variable d'âge de fin d'études, notée *agef*, dans le fait d'être en emploi, alors l'équation du modèle logit s'écrit :

$$P(y_i = 1|agef_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 agef_i}}$$

I.2.b Les variables binaires

Deuxième cas, celui d'une variable binaire. Reprenons l'exemple ci-dessus mais en remplaçant l'âge par le sexe. Pour l'introduire dans le modèle, on transforme d'abord cette variable de sexe en deux variables indicatrices, notées $\mathbb{1}(sexe_i = h)$ et $\mathbb{1}(sexe_i = f)$. La première (*resp.* seconde) vaut 1 si l'individu i est de sexe masculin (*resp.* féminin), et 0 sinon. On a, pour chaque individu i, $\mathbb{1}(sexe_i = h) + \mathbb{1}(sexe_i = f) = 1$.

Introduisons maintenant ces deux variables indicatrices dans le modèle :

$$P(y_i = 1|sexe_i) = \frac{1}{1 + e^{-\beta_0 - \beta_1 \mathbb{1}(sexe_i = h) - \beta_2 \mathbb{1}(sexe_i = f)}} \tag{17}$$

Puisque $\mathbb{1}(sexe_i = h) + \mathbb{1}(sexe_i = f) = 1$, on peut écrire :

$$\begin{aligned}&\beta_0 + \beta_1 \mathbb{1}(sexe_i = h) + \beta_2 \mathbb{1}(sexe_i = f)\\ &\quad = (\beta_0 + c) + (\beta_1 - c)\mathbb{1}(sexe_i = h) + (\beta_2 - c)\mathbb{1}(sexe_i = f)\end{aligned}$$

$$= \tilde{\beta}_0 + \tilde{\beta}_1 \mathbb{1}(sexe_i = h) + \tilde{\beta}_2 \mathbb{1}(sexe_i = f)$$

où $\tilde{\beta}_0 = \beta_0 + c$, $\tilde{\beta}_1 = \beta_1 - c$ et $\tilde{\beta}_2 = \beta_2 - c$, avec c constante quelconque pouvant prendre n'importe quelle valeur. L'équation (17) s'écrit donc aussi :

$$P(y_i = 1|sexe_i) = \frac{1}{1 + e^{-\tilde{\beta}_0 - \tilde{\beta}_1 \mathbb{1}(sexe_i = h) - \tilde{\beta}_2 \mathbb{1}(sexe_i = f)}}$$

Il y a donc une infinité de jeux de paramètres – donc d'équations – conduisant au même modèle. On dit que les paramètres du modèle ne sont pas *identifiés*. Or le modèle doit être représenté par une équation et une seule, c'est-à-dire par un jeu de paramètres et un seul.

Pour ce faire, on choisit une modalité qui fera office de référence, et on force à zéro le paramètre β correspondant. Par exemple, si on retient *homme* comme modalité de *référence* de la variable sexe, le paramètre associé – β_1 – est forcé à 0. La variable indicatrice $\mathbb{1}(sexe_i = homme)$ disparaît du modèle. La situation de la femme, mesurée par le paramètre $\tilde{\beta}_2$, est évaluée en référence à celle de l'homme.

I.2.c Les variables polytomiques

La démarche est la même pour une variable polytomique. On commence par la transformer en autant de variables indicatrices qu'il y a de modalités. On les introduit toutes dans le modèle, sauf une, pour les mêmes raisons que précédemment : pour obtenir l'unicité du jeu des paramètres associés à la variable, on en annule un. Cela signifie qu'on exclut du modèle une des indicatrices. La modalité qu'elle représente est appelée *modalité de référence*. Une variable polytomique à M modalités est donc remplacée par $M - 1$ indicatrices.

Supposons qu'on ait choisi la modalité $m = 1$ comme référence et qu'aux modalités 2, 3, ..., M soient associés respectivement les paramètres β_2, β_3, ..., β_M. Alors la situation d'un individu dans l'état m ($1 < m \leq M$), mesurée par le paramètre β_m associé à (l'indicatrice représentant) la modalité m, est évaluée *en référence à* la situation d'un individu dans l'état 1. En d'autres termes, la valeur d'un paramètre β_m est *relative*. Il s'ensuit que le choix de la modalité de référence d'une variable polytomique a une incidence sur les valeurs des paramètres associés aux autres modalités de la variable, mais n'influe pas sur l'écart entre deux paramètres quelconques. En effet, soit m_1, m_2 et m_3 trois modalités de la variable. On a :

$$\beta_{m_3} - \beta_{m_2} = (\beta_{m_3} - \beta_{m_1}) - (\beta_{m_2} - \beta_{m_1})$$

La différence entre β_{m_3} et β_{m_2} est la même, que la modalité de référence soit m_1 (auquel cas $\beta_{m_1} = 0$) ou une autre (auquel cas $\beta_{m_1} \neq 0$, sans conséquence sur la différence $\beta_{m_3} - \beta_{m_2}$).

Pour choisir la modalité de référence, on tiendra compte de plusieurs aspects. Elle doit d'abord recueillir un nombre suffisant d'observations pour donner de la robustesse aux estimations des paramètres (et pour qu'elle mérite son appellation : une

modalité rare ne peut faire référence ...). Le choix doit aussi pouvoir faciliter les commentaires des résultats. Dans le cas d'une variable ordonnée, par exemple, on prendra en général comme référence la modalité la plus « faible », à condition qu'elle recueille un nombre suffisant d'observations, l'idée étant qu'une variable qualitative ordonnée est un peu comme une variable numérique, dont les valeurs sont par définition ordonnées. Dans le cas d'une variable non ordonnée, s'il n'y a pas de choix évident, on pourra retenir la modalité modale, celle qui rassemble le plus d'individus, ou bien celle pour laquelle la répartition des individus entre les catégories C_0 et C_1 est proche de celle constatée sur l'ensemble de l'échantillon.

Mentionnons enfin qu'une variable numérique peut être transformée en variable polytomique ordonnée et traitée comme telle. Par exemple, l'âge peut être « découpé » en trois tranches – moins de 35 ans, de 35 à 45 ans, plus de 45 ans. À ces trois tranches sont associées trois variables indicatrices ($\mathbb{1}(age < 35)$, $\mathbb{1}(35 \leq age < 45)$, $\mathbb{1}(45 \leq age)$). Si on souhaite mettre en évidence des effets non-linéaires de l'âge sur la probabilité d'appartenance, le fait qu'elle soit par exemple plus élevée pour les deux tranches d'âge extrêmes, on a intérêt à retenir la modalité intermédiaire comme référence, c'est-à-dire exclure l'indicatrice $\mathbb{1}(35 \leq age < 45)$. On s'attend à ce que les paramètres associés aux deux autres indicatrices soient positifs.

I.3 Estimation des paramètres du modèle

I.3.a La méthode du maximum de vraisemblance

Pour estimer les paramètres du modèle, on utilise la méthode du maximum de vraisemblance. Pour expliquer en quoi elle consiste, nous allons partir d'un exemple simplifié à l'extrême.

On observe un échantillon de trois individus tirés, dans une population d'intérêt, aléatoirement et indépendamment les uns des autres. On connaît de ces individus une seule caractéristique, notée x_1, dont les valeurs sont respectivement $x_{11} = 2$, $x_{12} = 1$ et $x_{13} = 3$. Soit y la variable binaire repérant la catégorie d'appartenance. Dans cet échantillon, on observe que le premier individu appartient à la catégorie 1 ($y_1 = 1$), le deuxième à la catégorie 0 ($y_2 = 0$) et le troisième à la catégorie 1 ($y_3 = 1$).

La probabilité d'observer cet échantillon est celle d'observer conjointement $y_1 = 1$ compte tenu de la valeur de x_{11}, $y_2 = 0$ compte tenu que $x_{12} = 1$ et $y_3 = 1$ sachant que $x_{13} = 3$. Cette probabilité s'écrit :

$$\mathcal{P} = P(y_1 = 1|x_{11}, y_2 = 0|x_{12}, y_3 = 1|x_{13})$$

Puisque les individus ont été tirés indépendamment les uns des autres, cette probabilité est égale au produit des trois probabilités individuelles :

$$\mathcal{P} = P(y_1 = 1|x_{11})\, P(y_2 = 0|x_{12})\, P(y_3 = 1|x_{13})$$

En remplaçant les probabilités individuelles par leurs expressions (5) [15], la probabilité $\mathcal{P}$ d'observer l'échantillon tiré s'écrit :

$$\begin{aligned}\mathcal{P} &= \frac{1}{1+\exp[-\beta_0 - \beta_1 x_{11}]} \frac{\exp[-\beta_0 - \beta_1 x_{12}]}{1+\exp[-\beta_0 - \beta_1 x_{12}]} \frac{1}{1+\exp[-\beta_0 - \beta_1 x_{13}]} \\ &= \frac{1}{1+\exp[-\beta_0 - \beta_1.2]} \frac{\exp[-\beta_0 - \beta_1.1]}{1+\exp[-\beta_0 - \beta_1.1]} \frac{1}{1+\exp[-\beta_0 - \beta_1.3]} \end{aligned} \tag{18}$$

La probabilité $\mathcal{P}$ dépend des deux quantités (paramètres) β_0 et de β_1, inconnues à ce stade. Elle peut être plus ou moins élevée selon les valeurs de β_0 et de β_1. Il s'agit de les déterminer de manière unique. Pour le faire, on raisonne comme suit.

Lorsqu'on tire un échantillon de trois individus tels que $x_1 = 2$ pour le premier, $x_1 = 1$ pour le deuxième et $x_1 = 3$ pour le troisième, on a *a priori* 8 combinaisons possibles pour le triplet (y_1, y_2, y_3) : $(0,0,0)$, $(0,0,1)$, $(0,1,0)$, $(0,1,1)$,.... C'est la combinaison $(1,0,1)$ qu'on observe. L'idée de la méthode repose sur une quasi pétition de principe : si on observe effectivement cette combinaison, c'est parce qu'elle correspond à celle qu'on avait le plus de chances d'observer parmi les huit possibles, c'est la combinaison qui était la plus probable, la plus *vraisemblable* à observer.

En conséquence, les valeurs de β_0 et de β_1 à retenir sont celles qui rendent la plus

15. Pour l'individu n° 2, on a : $P(y_2 = 0|x_{12}) = 1 - P(y_2 = 1|x_{12})$.

élevée possible la probabilité $\mathcal{P}$ d'observer l'échantillon tiré, qui maximisent $\mathcal{P}$. D'où le nom de la méthode, étant entendu que la probabilité $\mathcal{P}$ est traditionnellement appelée *vraisemblance* du modèle. Les valeurs des deux paramètres sont donc celles qui annulent les deux dérivées partielles de $\mathcal{P}$:

$$\frac{\partial \mathcal{P}}{\partial \beta_0} = 0 \qquad \text{et} \qquad \frac{\partial \mathcal{P}}{\partial \beta_1} = 0$$

On montre que la fonction $\mathcal{P}$ est concave, ce qui fait que le point d'annulation des deux dérivées partielles correspond bien à un maximum.

Avec la forme (18) de la vraisemblance $\mathcal{P}$, les expressions des dérivées partielles sont assez compliquées. Pour simplifier, au lieu de maximiser $\mathcal{P}$, on maximise son logarithme $\ln \mathcal{P}$, ce qui revient au même puisque la fonction logarithme est strictement croissante. La quantité $\ln \mathcal{P}$ est appelée *log-vraisemblance*. Elle est la somme de trois logarithmes. Chaque dérivée partielle $\partial \ln \mathcal{P} / \partial \beta_0$ et $\partial \ln \mathcal{P} / \partial \beta_1$ est alors la somme de trois dérivées partielles (relativement) simples.

La généralisation à un échantillon de n individus est immédiate. Le raisonnement est exactement le même. On remarquera que, quelle que soit la valeur prise par la variable binaire y pour l'individu i, le logarithme de la probabilité individuelle de i, que i appartienne à C_0 ou à C_1, s'écrit toujours :

$$\ln P_i = y_i \ln \frac{1}{1 + e^{-x_i \beta}} + (1 - y_i) \ln \frac{e^{-x_i \beta}}{1 + e^{-x_i \beta}}$$

Avec cette notation, la log-vraisemblance pour l'échantillon des n individus est égale à :

$$\ln \mathcal{P} = \sum_{i=1}^{n} \left[y_i \ln \frac{1}{1 + e^{-x_i \beta}} + (1 - y_i) \ln \frac{e^{-x_i \beta}}{1 + e^{-x_i \beta}} \right] \tag{19}$$

Les valeurs des paramètres sont les solutions du système à $K + 1$ équations (il y a autant d'équations que de paramètres à estimer) :

$$\frac{\partial \ln \mathcal{P}}{\partial \beta} = 0 \tag{20}$$

où $\mathcal{P}$ est donnée par l'expression (19).

En théorie, de ce système de $K + 1$ équations à $K + 1$ inconnues β_k, on devrait déduire les β_k comme fonctions des variables x_k (pour $k = 0, \ldots, K$) [16]. Pour obtenir les valeurs estimées des β_k, traditionnellement notées $\hat{\beta}_k$, il suffirait alors de donner aux variables les valeurs qu'elles prennent dans l'échantillon. Le problème est qu'on ne peut pas procéder comme cela, car il n'y a pas d'expression analytique des β_k comme fonctions des variables composant le vecteur x. On est obligé d'utiliser des algorithmes qui recherchent pas à pas les valeurs des paramètres. Un des algorithmes

16. Ces fonctions sont appelées estimateurs des paramètres du modèle.

les plus souvent utilisés est celui de Newton-Raphson. Très schématiquement, il se déroule de la manière suivante. On part de valeurs initiales des $K+1$ paramètres du modèle (par exemple, $\beta_k = 0 \quad \forall\ k = 0, \ldots, K$). Puis on remplace chaque équation de (20) par son approximation linéaire autour de ces valeurs initiales. On résout le système ainsi formé et on obtient un premier jeu de valeurs des paramètres. On répète l'opération en remplaçant chaque équation de (20) par son approximation linéaire autour de ce premier jeu de paramètres. On résout le système ainsi formé, et ainsi de suite jusqu'à ce que les valeurs des paramètres ainsi déterminées ne changent (quasiment) pas lorsqu'on itère l'opération. On arrête alors la recherche des valeurs des β_k et les dernières obtenues sont les valeurs estimées des paramètres.

Grâce à elles, on peut calculer pour chaque individu i la probabilité d'appartenance à la catégorie C_1 prédite par le modèle, que l'on note $\hat{P}_i = \hat{P}(y_i = 1|x_i)$. Il suffit de remplacer, dans l'expression (5), β par $\hat{\beta}$, vecteur (colonne) des paramètres estimés $\hat{\beta}_k$ [17].

I.3.b Les propriétés des valeurs estimées des paramètres

Nous avons donc obtenu un jeu de valeurs estimées des paramètres du modèle, sur notre échantillon de n individus. Si nous avions tiré, dans notre population d'intérêt de N individus, un autre échantillon de même taille n, et si nous avions estimé les paramètres du modèle sur ce deuxième échantillon, nous aurions obtenu des valeurs des paramètres différentes des premières. Avec un autre échantillon de taille n, nous aurions obtenu encore d'autres valeurs. Et ainsi de suite. Par conséquent, les $K+1$ valeurs $\hat{\beta}_k$ estimées sur notre échantillon ne constituent qu'un ensemble de valeurs parmi toutes celles qu'on obtiendrait en estimant le modèle sur tous les échantillons possibles de n individus.

On montre que toutes ces valeurs possibles des paramètres estimés du modèle sont distribuées autour de la « vraie » valeur de β (i.e. des « vraies » valeurs des paramètres β_k associés aux x_k) selon approximativement une loi normale (de dimension $K+1$), ceci à condition que n soit suffisamment grand. En d'autres termes, les valeurs $\hat{\beta}_k$ estimées sur notre échantillon peuvent être considérées comme tirées dans une loi normale centrée autour de la « vraie » valeur de β. On dit que la loi de distribution asymptotique de l'estimateur des paramètres du modèle est la loi normale, dont la moyenne est la « vraie » valeur de β et la variance [18], inconnue, peut être estimée par des fonctions impliquant les valeurs estimées des paramètres et les valeurs prises sur l'échantillon par les variables x. On récupère ainsi les valeurs $\hat{\sigma}_k$ des écarts-types des $\hat{\beta}_k$.

Cela étant, la loi normale est distribuée de telle manière que 95% des valeurs

17. Pour éviter des lourdeurs d'écriture, on utilise la même notation – $\hat{\beta}_k$ – pour représenter à la fois l'estimateur de β_k (i.e. la fonction des variables x_k issue du système d'équations (20)) et la valeur estimée de β_k, c'est-à-dire celle prise par l'estimateur pour les valeurs des variables x_k observées sur l'échantillon d'étude.

18. Il s'agit plus précisément de la matrice de variance-covariance, de dimension $(K+1) \times (K+1)$. Les racines carrées des éléments diagonaux sont les écarts-types des paramètres estimés.

possibles des paramètres associés aux variables x_k sont comprises entre les valeurs $\beta_k - 1.96\hat{\sigma}_k$ et $\beta_k + 1.96\hat{\sigma}_k$. Toujours d'après les propriétés de la loi normale, 99% de ces valeurs possibles sont comprises entre $\beta_k - 2.58\hat{\sigma}_k$ et $\beta_k + 2.58\hat{\sigma}_k$. On a donc 95% de chances d'avoir :

$$\beta_k - 1.96\hat{\sigma}_k \leq \hat{\beta}_k \leq \beta_k + 1.96\hat{\sigma}_k$$

et 99% de chances que :

$$\beta_k - 2.58\hat{\sigma}_k \leq \hat{\beta}_k \leq \beta_k + 2.58\hat{\sigma}_k$$

Ces deux inégalités se réécrivent respectivement :

$$\hat{\beta}_k - 1.96\hat{\sigma}_k \leq \beta_k \leq \hat{\beta}_k + 1.96\hat{\sigma}_k \quad \text{et} \quad \hat{\beta}_k - 2.58\hat{\sigma}_k \leq \beta_k \leq \hat{\beta}_k + 2.58\hat{\sigma}_k$$

Autrement dit, on a 95% de chances que la « vraie » valeur de β_k soit dans l'intervalle :

$$I_{95} = [\hat{\beta}_k - 1.96\hat{\sigma}_k, \hat{\beta}_k + 1.96\hat{\sigma}_k]$$

et 99% de chances qu'elle appartienne à l'intervalle :

$$I_{99} = [\hat{\beta}_k - 2.58\hat{\sigma}_k, \hat{\beta}_k + 2.58\hat{\sigma}_k]$$

L'intervalle I_{95} est appelé *intervalle de confiance à 95%* du paramètre estimé $\hat{\beta}_k$. L'intervalle I_{99} est l'intervalle de confiance à 99% de $\hat{\beta}_k$.

On dit qu'un paramètre est estimé *avec précision* lorsque son écart-type (estimé) est faible. Dans ce cas, l'intervalle de confiance est peu étendu. La vraie valeur du paramètre est selon toute probabilité peu éloignée de la valeur estimée $\hat{\beta}_k$. A l'inverse, une estimation imprécise se manifeste par un écart-type important. C'est ce qui peut se produire avec une variable mal mesurée. Un exemple typique est celui du « revenu annuel total du foyer » lorsqu'il est renseigné par le ménage enquêté. Il est souvent arrondi à la centaine d'euros. La valeur répondue peut être assez éloignée de la vraie valeur, si bien que les valeurs du paramètre associé qu'on obtiendrait sur les différents échantillons de taille n risquent d'être plus dispersées (l'écart-type du paramètre estimé est donc plus important) que celles qu'on aurait si la variable était mieux renseignée.

Un dernier point. Si n est proche de N, alors les différents échantillons de taille n tirés dans la population d'intérêt auront un ensemble commun d'individus assez important. Les valeurs des paramètres estimées sur ces différents échantillons seront alors (très) proches les unes des autres puisqu'elles auront été déterminées en très large partie sur les mêmes individus. Ceci explique pourquoi les écarts-types des paramètres estimés sont plus faibles lorsque la taille n de l'échantillon est importante. Autrement dit, la précision des estimations augmente avec la taille de l'échantillon.

I.4 Les indicateurs de qualité du modèle estimé

Il y a deux manières d'évaluer la qualité globale du modèle estimé. La première s'appuie sur sa vraisemblance, la seconde sur les probabilités $\hat{P}_i$ prédites par le modèle.

I.4.a Les indicateurs fondés sur la vraisemblance du modèle

Le premier type d'indicateurs de qualité du modèle découle de la question suivante. Par définition, la vraisemblance du modèle est maximale pour les valeurs $\hat{\beta}_k$ des paramètres, mais est-elle suffisamment élevée pour que l'on considère le modèle ainsi estimé comme un « bon modèle » ? Pour le savoir, il faut d'abord régler deux points préalables. En premier lieu, la valeur absolue de la vraisemblance n'a pas de sens en soi. Il faut la comparer à une référence. Celle qui est généralement retenue est la vraisemblance du modèle qu'on peut considérer comme le plus « pauvre », celui sans aucune variable explicative, hormis le terme constant. C'est donc l'écart ou le rapport de la vraisemblance L [19] du modèle estimé à la vraisemblance du modèle sans variable explicative, notée L_0, qui importe. Deuxième point, il faut trouver une expression adéquate d'un indicateur qui fasse intervenir l'écart ou le rapport des deux vraisemblances et qui rende bien compte de la qualité du modèle.

Un des premiers indicateurs que l'on trouve dans la littérature est dû à McFadden. Il est noté ρ^2 et parfois appelé pseudo-R^2 (de McFadden). Il s'écrit :

$$\rho^2 = 1 - \frac{\ln L}{\ln L_0}$$

La log-vraisemblance du modèle est la somme de n quantités qui sont toutes négatives puisque chacune d'elles est le logarithme d'une probabilité, qui est inférieure à 1 par définition. Par conséquent, $\ln L < 0$. Comme le modèle avec variable explicative est plus « vraisemblable » que le modèle sans variable explicative, on a $\ln L_0 < \ln L < 0$. En conséquence, on a bien $0 < \rho^2 < 1$, et l'indicateur ρ^2 augmente avec la (log)vraisemblance $\ln L$ du modèle.

Pour tenir compte du nombre de variables introduites dans le modèle, Ben-Akiva et Lerman [20] ont proposé l'indicateur $\bar{\rho}^2$, qui « ajuste » le ρ^2 :

$$\bar{\rho}^2 = 1 - \frac{\ln L - (K+1)}{\ln L_0} \tag{21}$$

où K est le nombre de variables (hormis le terme constant). Attention ! Une variable catégorielle à m modalités compte pour n$m-1$ variables.

Les propriétés de ρ^2 – le fait qu'il soit compris entre 0 et 1 et qu'il augmente avec la

19. La vraisemblance est traditionnellement notée L à cause de la dénomination anglo-saxonne de la vraisemblance : *Likelihood*.

20. M. Ben-Akiva and S. Lerman, *Discrete Choice Analysis : Theory and Application to Travel Demand*, MIT Press, 1985.

qualité (la vraisemblance) du modèle – font penser au coefficient de détermination R^2 d'un modèle de régression linéaire classique. Mais il n'en possède pas toutes les propriétés. Notamment, ses valeurs ne couvrent pas tout l'intervalle [0,1], elles restent faibles même lorsqu'un modèle est considéré comme « très bon ». Estrella [21] a proposé un autre indicateur qui pallie ces défauts :

$$\phi_0 = 1 - \left(\frac{\ln L}{\ln L_0} \right)^{-\frac{2}{n} \ln L_0}$$

Il existe un autre type d'indicateurs, toujours fondés sur la vraisemblance du modèle. Ils sont appelés *critères d'information*. Cette dénomination provient de ce qu'ils mesurent la perte d'information due au fait que l'on remplace la réalité par un modèle. De ce point de vue, plus la valeur du critère est faible, plus la perte d'information est limitée, et donc meilleur est le modèle. Ces indicateurs sont des outils permettant de départager plusieurs modèles concurrents qui reposent sur des variables x différentes, à condition qu'ils soient tous estimés sur le même ensemble de données.

Les deux critères les plus utilisés sont le critère d'Akaïke (AIC) et le critère de Schwartz (SC). Ils s'écrivent respectivement :

$$AIC = 2(K+1) - 2\ln(L) \tag{22}$$

et :

$$SC = (K+1)\ln(n) - 2\ln(L) \tag{23}$$

où L est la valeur maximale de la vraisemblance (i.e. la valeur de la vraisemblance calculée avec les valeurs estimées des paramètres) et K le nombre de variables du modèle.

La présence de K dans les expressions (22) ou (23) se justifie. D'une manière générale, la vraisemblance maximale L d'un modèle peut être augmentée par le seul fait d'ajouter des variables (donc des paramètres) supplémentaires, quelle qu'en soit la pertinence. Dans ces conditions, on peut artificiellement augmenter la « qualité » d'un modèle en ajoutant n'importe quelle caractéristique individuelle dans la liste des x. La présence de K permet alors d'empêcher ce travers en pénalisant l'ajout de variables : avec des variables supplémentaires $\ln(L)$, certes, augmente, mais $2K$ ou $K\ln(n)$ aussi, et on ne sait pas *a priori* lequel des deux l'emporte. On ne sait pas si AIC ou SC augmente ou diminue. Sur cet aspect des choses, le critère de Schwartz pénalise plus fortement l'ajout de variables que ne le fait le critère d'Akaïke.

Parce qu'ils font intervenir de manière antagoniste la vraisemblance et le nombre de variables, ces indicateurs soulignent une qualité que doit avoir un modèle : la parci-

21. A. Estrella, « A New Measure of Fit for Equations With Dichotomous Dependent Variables », *Journal of Business & Economic Statistics*, 1998, vol. 16, n° 2.

monie. On doit veiller à cela surtout lorsqu'on introduit dans le modèle des variables polytomiques. En effet, la prise en compte d'une variable à, mettons, 10 modalités introduit 9 variables indicatrices supplémentaires. La qualité du modèle, mesurée par un critère d'information, risque d'en être affectée.

I.4.b Les indicateurs fondés sur les prédictions du modèle

Une autre manière d'évaluer la qualité d'un modèle est de regarder s'il reproduit correctement la réalité, si la catégorie d'appartenance qu'il prédit pour chaque individu i correspond bien à la catégorie à laquelle i appartient effectivement, si, en d'autres termes, le modèle s'ajuste bien à la réalité. D'où l'appellation d'*indicateur d'ajustement* parfois utilisée. Cette idée, à première vue naturelle, est difficile à mettre en œuvre, car on est amené à comparer des grandeurs qui ne sont pas de même nature. En effet, ce que l'on observe pour chaque individu i de l'échantillon, est la variable d'appartenance catégorielle y_i, qui vaut 0 si $i \in C_0$ ou 1 si $i \in C_1$. Ce que prédit le modèle est une probabilité d'appartenance à C_1, notée $\hat{P}_i$ pour l'individu i (voir section I.3.a), qui varie entre 0 et 1. Pour comparer prédictions et réalisations, il faut donc (essayer de) calculer l'appartenance (et non la probabilité d'appartenance) prédite par le modèle, qu'on notera $\hat{y}_i$. Malheureusement, on va le voir, il n'existe pas de solution entièrement satisfaisante.

Une première solution est d'adopter la règle de décision suivante. Si $\hat{P}_i$ est supérieure à 0.5, alors la catégorie d'appartenance prédite est C_1 : $\hat{y}_i = 1$. Dans le cas contraire, la catégorie prédite est C_0 : $\hat{y}_i = 0$. Il suffit alors de compter le nombre de fois où y_i et $\hat{y}_i$ coïncident. Mais ce premier comptage est insuffisant. Pour le voir, supposons que dans un échantillon de $n = 100$ personnes, on en observe 80 en catégorie C_0 et 20 en catégorie C_1 [22]. Supposons que le modèle prédise correctement l'appartenance de 70 de ces 80 individus, c'est-à-dire qu'on ait $\hat{y}_i = 0$ dans 87.5% (70/80) des cas. Même si aucun des cas d'appartenance à C_1 n'est correctement prédit, la part de bonnes prédictions par le modèle est tout de même de 70%. Pourtant, on ne peut pas soutenir qu'un modèle qui « rate » *tous* les cas d'appartenance à C_1 soit un bon modèle. Il faut donc compter le nombre de fois où on a $\hat{y}_i = y_i = 0$ *et* le nombre de fois où $\hat{y}_i = y_i = 1$, c'est-à-dire le nombre de *paires concordantes*.

Cette manière de faire ne règle pas tous les cas de figure, en particulier ceux où il y a un fort déséquilibre dans la répartition observée des individus entre les deux catégories. Supposons, en effet, que 95% soient en C_0 et 5% en C_1. Si on garde la règle de décision qui veut que l'appartenance prédite $\hat{y}_i$ soit égale à 1 si la probabilité prédite dépasse le seuil de 0.5, alors étant donné la rareté des cas observés d'appartenance à C_1, il est fort possible qu'aucun des $\hat{P}_i$ calculés ne dépasse 0.5. La solution serait alors de retenir comme seuil non pas 0.5 mais 0.05, conformément à la répartition globale des individus entre C_0 et C_1. Mais on risque cette fois-ci de ne pas prédire assez d'individus de catégorie C_0. Par conséquent, le « bon » seuil se situe

22. L'exemple est tiré de J.M. Wooldridge, *Introductory Econometrics. A Modern Approach*, South-Western, 4th ed., 2009.

probablement entre 0.05 et 0.5. Il n'y a pas de règle évidente pour le déterminer.

Quoi qu'il en soit, les paires concordantes et discordantes entrent dans le calcul d'indicateurs de qualité prédictive du modèle. Un des plus utilisés est le *Somers' D* qui correspond à l'écart en valeur absolue (et divisé par 100) entre le pourcentage de paires concordantes et le pourcentage de paires discordantes. Tant que les catégories C_0 et C_1 ne sont pas trop déséquilibrées, cet indicateur est somme toute valide.

Wooldridge [23] a proposé un indicateur très intéressant, qui est dans l'esprit du coefficient R^2 de détermination d'un modèle linéaire classique. Il l'a de fait transposé au cas du modèle binaire, i.e. où la variable y est binaire. On peut donc le dénommer pseudo-R^2 (de Wooldridge). Wooldridge rappelle d'abord la propriété selon laquelle, dans le cas du modèle classique, le R^2 est égal au carré de la corrélation empirique des y_i et des $\hat{y}_i$, et la prédiction de y n'est autre que l'espérance conditionnelle (estimée) de y : $E(y|x)$. Dans le cas binaire où y prend les valeurs 1 ou 0, on a :

$$E(y|x) = 1.P(y = 1|x) + 0.P(y = 0|x) = P(y = 1|x)$$

Par conséquent :

$$\hat{y}_i = \hat{P}(y_i = 1|x_i)$$

Il s'ensuit que le pseudo-R^2 proposé par Wooldridge (et noté ici pR^2) est égal au carré du coefficient de corrélation des y_i et des $\hat{P}(y_i = 1|x_i)$, quantité que l'on peut calculer une fois connues les valeurs estimées des paramètres β. On peut aussi calculer, par analogie au R^2 ajusté du modèle linéaire classique, un pseudo-R^2 ajusté, qui tient compte du nombre K de variables introduites dans le modèle :

$$p\bar{R}^2 = 1 - \frac{(n-1)(1-pR^2)}{(n-K-1)}$$

23. J.M. Wooldridge, *op. cité*, page 582.

I.5 Les tests sur les paramètres estimés : évaluation de leur significativité statistique

Les tests sur les paramètres estimés du modèle permettent de savoir si les variables associées influent sur l'affectation à l'une ou l'autre catégorie. Leur mise en oeuvre dépend de la nature des variables introduites dans l'analyse.

I.5.a Les paramètres des variables continues ou binaires

Comme on l'a vu précédemment (section I.2), une variable continue, par exemple la variable mesurant l'âge de la personne, est introduite telle quelle dans le modèle. Dans le cas d'une variable binaire, on définit deux indicatrices et on n'en retient qu'une. Par exemple, la variable renseignant sur le sexe de l'individu i est introduite sous la forme $\mathbb{1}(sexe_i = femme)$. Dans les deux cas de figure – continue ou binaire – la variable est associée à un paramètre et un seul [24].

Notons x_1 la variable binaire considérée et β_1 le paramètre associé. On se demande si x_1 joue un rôle, c'est-à-dire si le paramètre β_1 qui lui est associé est ou non différent de zéro : s'il est nul, la variable disparaît de l'expression (5) du modèle et ne joue donc plus aucun rôle. Pour le savoir, il faut réaliser un test d'hypothèse. Cette démarche passe par la définition de trois objets : l'hypothèse dite « nulle » (notée H_0), l'hypothèse dite « alternative » (notée H_a) et la statistique de test S.

L'hypothèse nulle est celle que l'on cherche à rejeter. Disant cela, il peut sembler *a priori* peu naturel que la démarche consiste à infirmer une hypothèse plutôt qu'à la valider. Cela tient à des raisons de nature statistique, qu'on verra plus loin.

L'objectif du modèle étant de mettre en évidence des traits distinctifs des deux catégories, on espère que les variables qui y sont introduites distinguent effectivement les deux populations, c'est-à-dire que les paramètres associés ne sont pas nuls. Pour la variable x_1, l'hypothèse nulle à poser et qu'on espère rejeter est donc :

$$H_0 : \ \beta_1 = 0$$

Par contraste, l'hypothèse alternative est celle qu'on est prêt à accepter en cas de rejet de H_0. Elle s'écrit :

$$H_a : \ \beta_1 \neq 0$$

À ce stade, la démarche semble immédiate : il suffit de regarder si la valeur de β_1 estimée sur l'échantillon dont on dispose est nulle ou non. C'est évidemment plus compliqué que cela, car la valeur $\hat{\beta}_1$ est « incertaine ». On se rappelle en effet (voir la section I.3.b) que la valeur estimée est une valeur particulière parmi toutes celles qu'on aurait obtenues en estimant le paramètre sur d'autres échantillons de même taille n. Toutes ces valeurs potentielles sont distribuées selon une loi normale centrée

24. Pour être exact, rappelons que dans le cas d'une variable binaire, il y a aussi un paramètre associé à la modalité de référence, mais il est forcé à 0.

autour de la « vraie » valeur de β_1, valeur qu'on ne connaît pas.

Pour trancher entre H_0 et H_a, on a recours à la statistique de test, qui permet précisément de prendre en compte la distribution des valeurs potentielles du paramètre estimé. Dans notre cas particulier, la statistique de test est :

$$S = \frac{\hat{\beta}_1}{\hat{\sigma}_1}$$

où $\hat{\beta}_1$ et $\hat{\sigma}_1$ sont les estimateurs [25] respectivement de β_1 et de son écart-type σ_1.

On conduit le test d'hypothèse de la manière suivante. On suppose dans un premier temps que l'hypothèse H_0 est vraie. Sous cette hypothèse, il a été établi que $\hat{\beta}_1/\hat{\sigma}_1$ suit la loi normale centrée (i.e. dont la moyenne est nulle) et réduite (i.e. dont l'écart-type vaut 1) [26].

Les conclusions du test vont dépendre de deux facteurs. Le premier est la valeur prise par S sur l'échantillon à notre disposition. Le second est la part de risque qu'on est prêt à assumer en rejetant l'hypothèse nulle. Car, de par la nature probabiliste de S due au fait qu'on travaille sur un échantillon parmi d'autres possibles, on n'est jamais certain de prendre la bonne décision, on n'est pas à l'abri de l'erreur consistant à rejeter H_0 alors qu'*en réalité* le paramètre β_1 (qu'on n'observe pas, rappelons-le) est nul.

Cela posé, si H_0 est vraie, en vertu des propriétés de la loi normale, 90% des valeurs potentielles de $\hat{\beta}_1/\hat{\sigma}_1$ (i.e. celles qui seraient obtenues sur les différents échantillons de taille n) sont comprises entre (environ) -1.65 et + 1.65, 95% entre -1.96 et + 1.96, 99% entre -2.58 et + 2.58, ... Cela implique qu'il est peu probable *a priori* que la valeur de S estimée sur l'échantillon d'étude soit supérieure en valeur absolue à 1.96 : on a *a priori* 5% de chances, au maximum, que ce soit le cas.

Supposons qu'on ait, mettons, $S = 2$. Le fait qu'on tombe sur une valeur aussi peu probable (5% de chances au maximum qu'elle soit observée) peut alors faire douter de l'hypothèse de départ, qui postule que la « vraie » valeur β_1 soit nulle. Dans ces conditions, on est prêt à la remettre en cause. Mais il y a un risque que, ce faisant, on ne prenne pas la bonne décision. Ce risque correspond précisément au nombre de fois où la décision de remettre H_0 en cause est discutable, c'est-à-dire au nombre de fois où on peut observer une valeur jugée peu probable. Il est ici égal à 5%. Dans la littérature, ce risque est appelé *risque de première espèce*. On l'appelle aussi, beaucoup plus fréquemment, *seuil de significativité* et on dit alors que l'on rejette l'hypothèse nulle au seuil de 5% (i.e. avec un risque maximal de 5% de se tromper). On dit aussi que le paramètre est statistiquement significatif au seuil de 5%.

Bien sûr, plus la valeur observée de S est élevée (en valeur absolue), plus le risque

25. Voir la section (I.3.a) et la note 17 de la section.

26. En toute rigueur, la statistique de test suit une loi de Student. Mais lorsque l'échantillon est suffisamment important (i.e. n élevé, ce que nous supposons ici), cette loi est assimilable à la loi normale.

de se tromper en rejetant l'hypothèse nulle est faible. Ainsi, avec la valeur de 2.58, le risque est de 1% : le paramètre est statistiquement significatif au seuil de 1%. Plus généralement, on peut associer à toute valeur de S une probabilité de se tromper. Par exemple, la valeur de 2.2 correspond à une probabilité de 0.0278. Dans ce cas de figure, on peut rejeter H_0 au seuil de 3% mais pas au seuil de 2%. Cette valeur est appelée par les anglo-saxons *p-value*. Elle est automatiquement calculée par les logiciels courants de statistique, comme SAS.

Supposons maintenant qu'on trouve 1.8 comme valeur de la statistique de test. Si on fixe à 5% le seuil à partir duquel une valeur de la satistique peut être considérée comme peu probable, alors la valeur de 1.8 ne remet pas en cause la validité de H_0 puisqu'elle fait partie des 95% des valeurs considérées ici comme probables. En conséquence, sur cette base, on ne peut pas rejeter H_0.

Faut-il pour autant accepter H_0, c'est-à-dire affirmer que β_1 est nul, que x_1 n'a aucune influence ? En acceptant l'hypothèse nulle, on risque de se tromper si, en réalité, c'est l'hypothèse alternative qui est vraie. On risque d'accepter à tort l'hypothèse nulle. Dans la littérature, ce risque est appelé *risque de deuxième espèce*. La prudence commande de ne pas accepter l'hypothèse, de se borner à dire que *sur la base de l'échantillon à notre disposition*, on ne peut pas rejeter la possibilité que le paramètre soit nul. Il y a deux raisons à cela. D'abord, d'une manière générale, pour pouvoir calculer le risque de deuxième espèce du test d'hypothèse, il faut connaître la loi de probabilité de la statistique de test S sous l'hypothèse H_a, ce qui est très exceptionnellement possible. Ensuite, il faut se rappeler que la décision d'accepter H_0 pourrait être remise en cause si on disposait d'un échantillon de taille plus importante, avec lequel les estimations seraient plus précises, c'est-à-dire les valeurs potentielles de $\hat{\beta}_1$ plus resserrées (voir fin de la section I.3.b).

Replaçons-nous dans le cas où la valeur de la statistique de test est de 1.8, mais desserrons notre exigence sur le risque d'erreur si bien qu'on considère maintenant que les valeurs supérieures à 1.65 sont peu probables. Dans ce cas, on rejettera l'hypothèse nulle au seuil de 10%.

Deux dernières remarques. Dans le cas d'une variable binaire, le paramètre associé à la modalité de référence est fixé à 0 (section I.2). Dans ces conditions, tester la nullité de β_1, c'est tester la différence de deux situations. Par exemple, si on prend la modalité « homme » comme référence de la variable « sexe », β_1 est associé à l'indicatrice $\mathbb{1}(sexe_i = femme)$. Rejeter $\beta_1 = 0$, c'est dire que le sexe joue un rôle dans l'appartenance à C_0 ou à C_1.

Seconde remarque, d'autres statistiques de test que $\hat{\beta}_1/\hat{\sigma}_1$ sont possibles. La seule contrainte est qu'on en connaisse la loi de probabilité sous l'hypothèse nulle (i.e. lorsqu'on suppose que H_0 est vraie). Ainsi, la procédure de SAS qui estime les modèles logit utilise la statistique dite de Wald, $\hat{\beta}_1^2/\hat{\sigma}_1^2$, qui, sous l'hypothèse nulle, suit la loi du χ^2 à 1 degré de liberté. De par les propriétés de cette loi, le seuil de significativité de 10% correspond à la valeur 2.71 de la statistique de Wald, le seuil

de 5% à 3.84 et le seuil de 1% à 6.63.

I.5.b Les paramètres des variables polytomiques

On l'a vu (section I.2.c), une variable polytomique à M modalités est introduite dans le modèle sous la forme de $M-1$ variables indicatrices, la modalité non retenue faisant office de référence. Il y a donc, pour la variable, $M-1$ paramètres associés aux $M-1$ indicatrices. Supposons par exemple que la catégorie sociale, notée cs, soit codée en 4 postes : catégorie sociale dite très favorisée, catégorie favorisée, catégorie moyenne, et catégorie défavorisée. Dans ce cas, $M = 4$. Prenons comme référence la dernière nommée. La variable cs est donc représentée dans le modèle par trois indicatrices : $\mathbb{1}(cs_i = trfav)$, $\mathbb{1}(cs_i = fav)$ et $\mathbb{1}(cs_i = moy)$. Soit β_1, β_2 et β_3 les paramètres associés.

On peut tester la nullité de chacun de ces trois paramètres en suivant la même démarche que dans la section précédente. L'interprétation des résultats est la même que dans le cas d'une variable binaire. Par exemple, si on rejette ici l'hypothèse $\beta_3 = 0$, cela signifie que les individus de la catégorie sociale moyenne ne sont pas répartis entre les catégories C_0 et C_1 de la même manière que ceux faisant partie de la catégorie défavorisée (catégorie de référence pour laquelle, rappelons-le, $\beta_4 = 0$).

On peut aussi mettre en œuvre d'autres tests. Par exemple, regarder s'il y a une différence entre les catégories très favorisée et favorisée. Conformément à la démarche générale d'un test d'hypothèse (voir *supra*), l'hypothèse nulle est dans ce cas :

$$H_0 : \ \beta_1 = \beta_2$$

L'hypothèse alternative est :

$$H_a : \ \beta_1 \neq \beta_2$$

Avant d'expliciter la statistique de test, notons que H_0 et H_a peuvent s'écrire aussi :

$$H_0 : \ \beta_1 - \beta_2 = 0 \quad \text{et} \quad H_a : \beta_1 - \beta_2 \neq 0$$

En posant $\beta_{12} = \beta_1 - \beta_2$, on se retrouve dans le cas de figure de la section précédente, où on teste la nullité d'un paramètre, en l'occurrence β_{12}. La statistique de test est donc :

$$S = \frac{\hat{\beta}_{12}}{\hat{\sigma}_{12}}$$

Cette statistique suit, sous l'hypothèse nulle, la loi normale centrée réduite (i.e. de moyenne nulle et de variance unitaire). La difficulté est qu'on ne peut pas déduire l'écart-type estimé de β_{12} uniquement à partir des écarts-types estimés $\hat{\sigma}_1$ et $\hat{\sigma}_2$, car il faut faire aussi intervenir la corrélation entre $\hat{\beta}_1$ et $\hat{\beta}_2$. Heureusement, ce calcul est fait automatiquement dans les logiciels courants comme SAS.

Autre test possible : la nullité de l'ensemble des paramètres associés aux indicatrices représentant les modalités de la variable. Dans l'exemple de la catégorie sociale, l'hypothèse nulle s'écrit alors :

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = 0$$

Elle signifie que la catégorie sociale, du moins telle qu'elle est codée ici en quatre modalités, ne joue pas de rôle. L'hypothèse alternative est :

$$H_a : \ \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

Plus formellement, l'hypothèse nulle peut s'écrire aussi :

$$H_0 : \ Q\beta = 0$$

où Q est la matrice identité et β le vecteur-colonne des paramètres :

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

L'intérêt de l'écrire sous cette forme générale est qu'on peut l'appliquer à d'autres matrices Q et donc à d'autres tests sur les paramètres. La statistique de test s'écrit :

$$W = (Q\hat{\beta})'[Q(\hat{V}_{\hat{\beta}})Q']^{-1}(Q\hat{\beta})$$

où $\hat{V}_{\hat{\beta}}$ est la matrice de variance covariance estimée de $\hat{\beta}$. W suit une loi du χ^2 à q degrés de liberté, où q est le rang de la matrice Q. Les valeurs correspondant aux seuils de significativité de 10%, 5% ou 1% dépendent de q. Dans notre exemple où $q = 3$, les valeurs sont respectivement 6.25, 7.82 et 11.34. Dans la pratique, comme on le verra, il n'est pas nécessaire de calculer toutes ces quantités. En utilisant les instructions adéquates de la procédure SAS, on obtient directement la *p-value* du test.

Il faut signaler une difficulté dans ce type de test. Son résultat n'est pas toujours cohérent avec les résultats des tests menés sur chacune des variables indicatrices [27]. Il arrive qu'on puisse rejeter, au seuil de 5% par exemple, l'hypothèse H_0 de la nullité jointe des paramètres ($\beta_1 = \beta_2 = \beta_3 = 0$), alors qu'aucun des paramètres n'est statistiquement significatif à ce même seuil. L'inverse peut aussi se produire. En conséquence, il est prudent d'effectuer tous les tests et de confronter les résultats. Dans notre exemple, il y a donc quatre tests à réaliser, dont les hypothèses nulles

27. D. Le Blanc, S. Lollivier, M. Marpsat, D. Verger, « L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitative. Les modèles univariés à résidus logistiques et normaux », Document de travail n° 0001, Unité Méthodologie Statistique, INSEE.

sont $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$ et $\beta_1 = \beta_2 = \beta_3 = 0$.

I.6 Les valeurs des paramètres estimés : évaluation de leur significativité pratique

La significativité statistique permet d'évaluer le degré de certitude avec lequel on peut affirmer qu'une variable influe sur l'appartenance aux catégories C_0 ou C_1. Mais elle ne nous dit rien sur son importance. Cette information nous est fournie par ce que d'aucuns nomment la *significativité pratique*[28]. Ces deux notions doivent être clairement distinguées. Une variable peut avoir un impact important alors que le paramètre qui lui est associé est tout juste significatif au seuil de 5%. À l'inverse, le paramètre peut être significatif au seuil de 1% et la variable associée avoir un faible rôle dans l'appartenance à l'une ou l'autre catégorie.

La significativité pratique est mesurée par la valeur estimée du paramètre. Le problème est que cette valeur ne nous dit pas grand chose. Son signe en revanche donne une information immédiate. S'il est positif, alors la variable associée a un impact positif sur la probabilité d'appartenir à la catégorie C_1. Quant à savoir si l'impact est important ou non, on ne peut pas le deviner car le lien entre la probabilité d'appartenance à l'une des catégories et le paramètre – expression (5) ou (11) – est somme toute relativement complexe. Il faut faire appel à d'autres grandeurs statistiques : les *odds ratios* ou bien les *effets marginaux*.

I.6.a L'*odds ratio* en épidémiologie

Les résultats d'un logit sont souvent présentés sous la forme d'*odds ratio* – littéralement *rapport des cotes* (parfois appelé *rapport des chances* ou encore *rapport des risques relatifs*) – dont l'usage est traditionnel en épidémiologie. Il est en effet bien adapté à cette discipline, comme l'illustre l'exemple suivant.

Dans le but d'étudier l'influence de la consommation de tabac sur la survenance d'un cancer, une enquête a été conduite sur un échantillon de 300 personnes[29]. Les résultats donnent le tableau suivant :

	fumeur	non-fumeur
cancer	10	10
pas de cancer	90	190

Avec ces données, on peut calculer 4 probabilités. Par exemple, le risque pour un fumeur d'être atteint d'un cancer est égal à $10/(10+90) = 10\%$. Pour un non-fumeur, il vaut $10/(10+190) = 5\%$, soit deux fois moins.

Supposons que l'épidémiologiste ait eu un peu plus de temps et de moyens pour « recruter » un nombre plus important de personnes atteintes de cancer et qu'il dispose d'un échantillon de 120 patients. Si le « recrutement » a été fait de manière aléatoire, on devrait en principe observer, comme précédemment, une équirépartition

28. Cette dénomination est employée, entre autres, par J.M. Wooldridge, *op. cité*, page 135.

29. L'exemple est inspiré de : Emmanuel Lagarde, « Deux mesures d'association fréquemment utilisées en épidémiologie : l'Odds-Ratio et le Risque Relatif », *Transcriptases*, n° 72, mars 1999.

des fumeurs et des non fumeurs chez ces personnes atteintes de cancer. S'il y a toujours 280 personnes non atteintes, la distribution des patients est :

	fumeur	non-fumeur
cancer	60	60
pas de cancer	90	190

Avec ces données, le risque pour un fumeur d'être atteint d'un cancer est égal à $60/(60+90) = 40\%$. Pour un non-fumeur, il vaut $60/(60+190) = 24\%$. le rapport est maintenant inférieur à 2 (40% *vs* 24% ;40% *vs* 20% précédemment).

Cette mesure de l'impact du tabac sur la survenance d'un cancer – le rapport de ces deux risques – est insatisfaisante car elle dépend de la répartition, dans le plan de « recrutement », entre les personnes malades et les personnes saines. Pour éviter de tirer un échantillon dont la répartition malades/sains soit représentative de la population totale et conserver ainsi la souplesse de « recrutement », l'épidémiologiste a besoin d'une mesure du lien entre la consommation de tabac et la maladie qui soit invariante à la proportion : c'est l'*odds ratio*.

D'une manière générale, soit y la variable binaire mesurant la survenance d'un événement (exemple : être atteint d'un cancer) : $y = 1$ si l'événement survient, 0 sinon. Soit X une caractéristique binaire du patient (exemple : fumeur – $X = 1$ – *vs* non fumeur – $X = 0$). Les tables précédentes s'écrivent sous la forme :

	$X=1$	$X=0$
$y=1$	a	b
$y=0$	c	d

où a, b, c et d sont des effectifs. On appelle *cote* (au sens des parieurs) d'un événement le rapport de la probabilité de l'événement à celle de son complémentaire. On parle aussi de *risque relatif*. La cote peut se calculer pour chaque type de patient caractérisé par X. Pour les individus $X = 1$, il vaut :

$$\frac{p(y=1|X=1)}{p(y=0|X=1)} = \frac{\frac{a}{a+c}}{\frac{c}{a+c}} = \frac{a}{c}$$

Pour les individus $X = 0$, il vaut :

$$\frac{p(y=1|X=0)}{p(y=0|X=0)} = \frac{b}{d}$$

Le *rapport des cotes* (en anglais *odds ratio*) est le rapport de ces deux cotes. Il vaut donc :

$$OR = \frac{a}{c} / \frac{b}{d} \tag{24}$$

Ce rapport est invariant à la répartition entre patients malades ($y = 1$) et patients sains ($y = 0$). Si on prend, par exemple, k fois plus de $y = 1$, a et b sont remplacés par ka et kb (pour autant que le « tirage » des personnes malades soit aléatoire), mais OR ne change pas.

OR est bien une mesure d'association, qui mesure le lien entre la caractéristique X et la survenance de l'évenement $y = 1$. En effet, on a :

$$\begin{aligned} p(y = 1|X = 1) - p(y = 1|X = 0) &= \frac{a}{a+c} - \frac{b}{b+d} \\ &= \frac{bc}{(a+c)(b+d)}\,(OR - 1) \end{aligned}$$

En conséquence, si $OR = 1$, l'événement y et la caractéristique X sont indépendants. Si $OR > 1$ (resp. $OR < 1$), le lien entre y et X est positif (resp. négatif).

On insistera sur le fait que l'*odds ratio* n'est pas un rapport de probabilités, mais un rapport de rapports de probabilités. Il s'écrit :

$$OR = \frac{p(y = 1|X = 1)}{p(y = 0|X = 1)} \Big/ \frac{p(y = 1|X = 0)}{p(y = 0|X = 0)}$$

et non :

$$OR = p(y = 1|X = 1)/p(y = 1|X = 0)$$

Ainsi, avec le deuxième tableau de données, l'*odds ratio* est égal à $(60/90)/(60/190)$, c'est-à-dire $19/9$ soit $2,1$ environ. Le rapport des probabilités est, quant à lui, égal à $(60/150)/(60/250)$ soit $1,66$.

I.6.b *Odds ratio* et analyse multivariée

L'étude du lien entre tabac et cancer peut être affinée en introduisant d'autres variables comme l'âge pour savoir si, à âge fixé, l'impact de la consommation de tabac sur la survenance d'un cancer est toujours le même. Pour ce faire, les épidémiologistes ont naturellement recours au modèle logit, car le paramètre associé à la variable binaire « être ou non fumeur » s'interprète en termes d'odds ratio.

Pour le voir, on désigne par x_1 le fait de fumer ou non (β_1 étant le paramètre associé) et par $x_{(K-1)}$ les autres variables du modèle (avec $\beta_{(K-1)}$ comme paramètres associés). On part de l'égalité (5) ou (11) :

$$P(y = 1|x) = \frac{1}{1 + e^{-x\beta}}$$

On a aussi :

$$P(y = 0|x) = 1 - P(y = 1|x) = \frac{e^{-x\beta}}{1 + e^{-x\beta}}$$

Il vient alors :

$$\frac{P(y=1|x)}{P(y=0|x)} = e^{x\beta}$$

ou encore :

$$\ln\left[\frac{P(y=1|x)}{P(y=0|x)}\right] = x\beta \tag{25}$$

Fixons les variables $x_{(K-1)}$ à des valeurs quelconques $\tilde{x}_{(K-1)}$. Ecrivons l'expression (25) pour $x_1 = 1$ (fumeur) d'une part, et pour $x_1 = 0$ (non fumeur) d'autre part, les autres variables restant fixées à leurs valeurs $\tilde{x}_{(K-1)}$. On obtient respectivement :

$$\ln\left[\frac{P(y=1|x_1=1,\tilde{x}_{(K-1)})}{P(y=0|x_1=1,\tilde{x}_{(K-1)})}\right] = \beta_0 + \beta_1 + \beta_{(K-1)}\tilde{x}_{(K-1)}$$

et :

$$\ln\left[\frac{P(y=1|x_1=0,\tilde{x}_{(K-1)})}{P(y=0|x_1=0,\tilde{x}_{(K-1)})}\right] = \beta_0 + \beta_{(K-1)}\tilde{x}_{(K-1)}$$

Par différence, on obtient :

$$\ln\left[\frac{P(y=1|x_1=1,\tilde{x}_{(K-1)})}{P(y=0|x_1=1,\tilde{x}_{(K-1)})}\right] - \ln\left[\frac{P(y=1|x_1=0,\tilde{x}_{(K-1)})}{P(y=0|x_1=0,\tilde{x}_{(K-1)})}\right] = \beta_1$$

c'est-à-dire :

$$\ln\left[\frac{\frac{P(y=1|x_1=1,\tilde{x}_{(K-1)})}{P(y=0|x_1=1,\tilde{x}_{(K-1)})}}{\frac{P(y=1|x_1=0,\tilde{x}_{(K-1)})}{P(y=0|x_1=0,\tilde{x}_{(K-1)})}}\right] = \beta_1$$

En prenant l'exponentielle des deux membres de l'expression, on obtient :

$$\frac{\frac{P(y=1|x_1=1,\tilde{x}_{(K-1)})}{P(y=0|x_1=1,\tilde{x}_{(K-1)})}}{\frac{P(y=1|x_1=0,\tilde{x}_{(K-1)})}{P(y=0|x_1=0,\tilde{x}_{(K-1)})}} = \exp[\beta_1] \tag{26}$$

On reconnaît à gauche du signe d'égalité l'expression de l'odds ratio – cf (24) – associé à x_1, les autres variables observées étant fixées à des valeurs quelconques. L'*odds ratio* s'exprime donc très simplement en fonction du seul paramètre attaché à x_1. Sa valeur s'obtient en remplaçant β_1 par la valeur estimée $\hat{\beta}_1$.

La contrepartie de cette simplicité est la difficulté à exposer les résultats, c'est-à dire à traduire l'expression (26) en des termes aisément compréhensibles. La lecture précise de (26) consiste à dire qu'un fumeur (i.e. $x_1 = 1$) a $\exp[\hat{\beta}_1]$ fois plus de risques de développer un cancer (i.e. $y = 1$) qu'un non-fumeur ($x_1 = 0$), *en ayant en tête* que le risque est ici un *risque relatif* (voir page 40) mesuré par un rapport de probabilités et non par une simple probabilité. Le message n'est donc pas toujours aisé à faire passer, surtout si on vise un public non initié.

I.6.c Les effets marginaux

L'effet marginal [30] d'une variable est la seconde manière d'évaluer la significativité pratique du paramètre qui lui est associé. Cette seconde solution a l'avantage de rendre les résultats d'un *logit* plus faciles à lire qu'avec l'approche par les *odds ratio*. En revanche, l'effet marginal peut être estimé de plusieurs façons, qui ne conduisent pas exactement aux mêmes résultats.

Le calcul des effets marginaux dépend de la nature – discrète ou continue – de la variable. Commençons par le premier cas.

Prenons d'abord le cas d'une variable binaire, x_1 par exemple. Pour obtenir son effet marginal, on calcule la probabilité $P(y = 1|x)$ pour $x_1 = 1$ d'une part, et pour $x_1 = 0$ d'autre part. L'effet marginal de x_1 sur $P(y = 1|x)$ est la différence de ces deux probabilités :

$$\Delta = G(\beta_0 + \beta_1 + \beta_2 x_2 + \ldots + \beta_K x_K) - G(\beta_0 + \beta_2 x_2 + \ldots + \beta_K x_K) \tag{27}$$

où $G(x) = 1/[1 + e^{-x\beta}]$.

À la différence de l'*odds ratio* qui ne dépend que de β_1 – voir expression (26) – et peut donc être facilemement estimé en remplaçant β_1 par $\hat{\beta}_1$, l'effet marginal est fonction non seulement des paramètres du modèle mais aussi de toutes les variables x autres que x_1. Il faut donc leur attribuer des valeurs pour pouvoir estimer Δ.

Une première possibilité est de partir du niveau individuel, de calculer la quantité (27) pour chaque individu avec ses propres valeurs de x_2, x_3, …, x_K, en donnant aux β leurs valeurs estimées $\hat{\beta}$. On obtient ainsi la variation individuelle de $P(y = 1|x)$ due à la *seule* variation de x_1, c'est-à-dire en maintenant constantes les caractéristiques x_2, x_3, …, x_K de l'individu. L'effet marginal de x_1 est alors la moyenne des variations individuelles ainsi calculées. Formellement, on procède comme suit :

(i) On calcule pour chaque individu i de l'échantillon la valeur prédite de la quantité (27), i.e. celle obtenue en remplaçant β par $\hat{\beta}$:

$$\Delta_i = G(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_K x_{Ki}) - G(\hat{\beta}_0 + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_K x_{Ki})$$

(ii) On prend la moyenne arithmétique de ces n valeurs prédites.

L'effet marginal de x_1 sur $P(y = 1|x)$ est donc estimé par :

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{1 + e^{-(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 x_{2i}+\ldots+\hat{\beta}_K x_{Ki})}} - \frac{1}{1 + e^{-(\hat{\beta}_0+\hat{\beta}_2 x_{2i}+\ldots+\hat{\beta}_K x_{Ki})}}\right] \tag{28}$$

Une autre solution est de fixer les variables $x_2, \ldots, x_K$ à des valeurs quelconques : $\tilde{x}_2, \ldots, \tilde{x}_K$, les mêmes pour tous les individus. L'effet marginal de x_1 sur $P(y = 1|x)$

30. Certains auteurs – notamment J.M. Wooldridge (*op cité*, p 577) – parlent d'*effet partiel*. Nous avons choisi *marginal* plutôt que *partiel* car c'est le terme le plus souvent utilisé dans la littérature.

est alors estimé par :

$$\Delta = G(\beta_0 + \beta_1 + \beta_2\tilde{x}_2 + \ldots + \beta_K\tilde{x}_K) - G(\beta_0 + \beta_2\tilde{x}_2 + \ldots + \beta_K\tilde{x}_K)$$

On prend généralement comme valeurs $\tilde{x}_2, \ldots, \tilde{x}_K$, les moyennes des x : $\bar{x}_2, \ldots, \bar{x}_K$. Ce faisant, on se situe au « point moyen » de l'échantillon. Tout se passe alors comme si on calculait l'effet marginal de x_1 pour l'« individu moyen », qui est un individu fictif. Par exemple, si x_2 est la variable de sexe et si l'échantillon est composé à 60% d'hommes ($x_2 = 0$) et à 40% de femmes ($x_2 = 1$), alors $\bar{x}_2 = 0.4$ [31].

La première manière – expression (28) – de calculer l'effet marginal est en général retenue comme préférable à la seconde, car elle respecte mieux le caractère non linéaire de la relation entre les variables x et la probabilité $P(y = 1|x)$. La seconde manière considère implicitement que la moyenne des quantités $G(a_i)$, où $a_i = \beta_0 + \ldots + \beta_K x_K$, est égale à $G(\bar{a}_i)$, où $\bar{a}_i$ est la moyenne des a_i, ce qui n'est pas exact car la fonction G n'est pas linéaire.

Un cas particulier important, dont on reparlera dans la seconde partie du document (voir page 44) : le modèle à une seule variable, notée x_1, de nature binaire. L'effet marginal de x_1 est égal à $P(y = 1|x_1 = 1) - P(y = 1|x_1 = 0)$. La quantité $P(y = 1|x_1 = 1)$ (resp. $P(y = 1|x_1 = 0)$) s'estime par la proportion des individus appartenant à la catégorie C_1 ($y = 1$) parmi tous ceux dont $x_1 = 1$ (resp. $x_1 = 0$). Ces deux proportions sont directement calculables sur l'échantillon. Leur différence est exactement l'effet marginal de x_1.

Ces types de calcul s'étendent sans difficulté à une variable polytomique. Supposons, par exemple, que x_1, x_2 et x_3 représentent les trois modalités d'une variable polytomique. Prenons x_1 comme modalité de référence. On a donc $\beta_1 = 0$ (voir section I.2). De la même manière que l'effet marginal moyen d'une variable binaire est fondé sur l'expression (27), l'effet marginal moyen de la variable polytomique à trois modalités (i.e. les effets moyens des modalités de la variable) est fondé sur les deux quantités :

$$\begin{aligned}\Delta_2 = G(&\beta_0 + \beta_2 + \beta_4 x_4 + \ldots + \beta_K x_K) \\ &- G(\beta_0 + \beta_4 x_4 + \ldots + \beta_K x_K) \qquad (29a)\\ \Delta_3 = G(&\beta_0 + \beta_3 + \beta_4 x_4 + \ldots + \beta_K x_K) \\ &- G(\beta_0 + \beta_4 x_4 + \ldots + \beta_K x_K) \qquad (29b)\end{aligned}$$

Supposons maintenant la variable x_1 continue ou quasi-continue (comme l'âge). Cela a un sens de dériver l'expression (11) par rapport à x_1. On obtient, tous calculs

31. L'individu moyen est un hermaphrodite ! Notons que dans le cas d'une variable polytomique m modalités, on retient les valeurs moyennes des $m-1$ indicatrices représentant les $m-1$ modalités autres que la modalité de référence.

faits :

$$\frac{\partial P(y=1|x)}{\partial x_1} = \frac{\exp(x\beta)}{[1+\exp(x\beta)]^2}\beta_1$$

Cette quantité est l'effet marginal de x_1 sur $P(y=1|x)$. Par exemple, si x_1 est l'âge, l'impact sur la probabilité $P(y=1|x)$ du vieillissement d'un an (i.e. $\partial age = 1$) est égal à :

$$\frac{\exp(x\beta)}{[1+\exp(x\beta)]^2}\beta_1 \tag{30}$$

Comme dans le cas précédent, il y a deux possibilités pour calculer la valeur de l'effet marginal. La première consiste à calculer la quantité (30) pour chaque individu de l'échantillon, puis de prendre la moyenne arithmétique de ces quantités individuelles. L'effet marginal de x_1 sur $P(y=1|x)$ est estimé par :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\exp(x_i\hat{\beta})}{[1+\exp(x_i\hat{\beta})]^2}\hat{\beta}_1 = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\exp(x_i\hat{\beta})}{[1+\exp(x_i\hat{\beta})]^2}\right]\hat{\beta}_1 \tag{31}$$

La seconde solution est fixer les valeurs des autres variables (au « point moyen » de l'échantillon, par exemple). Là aussi, la première solution est préférable.

Deux remarques, pour terminer. Le calcul des effets marginaux d'une variable continue au moyen de (31) est pleinement justifié lorsque la variable se compte en unités de mesure, comme par exemple l'âge (mesuré en années) ou le revenu (mesuré en euros). En revanche, il l'est moins lorsque la variable n'a pas d'unité de mesure, comme une note reçue par un élève à une épreuve : mesurer l'effet d'une augmentation d'un point de la note sur la variable y n'a pas de valeur en soi puisque le correcteur peut choisir de noter sur 10 ou sur 20, ou utiliser un autre système de notation. En revanche, comme une note sert à classer les élèves, on peut les répartir en plusieurs groupes selon leur classement, par exemple 4 groupes selon les quartiles de la distribution des notes : le premier groupe comprendrait les 25% d'élèves les moins bien notés, le second groupe les 25% mieux notés que les précédents, et ainsi de suite. Cela revient à transformer la variable continue en variable polytomique à 4 modalités, et les effets marginaux se calculent comme indiqué précédemment.

Seconde remarque, les effets marginaux, même s'ils ont l'avantage de permettre une lecture plus immédiate que les *odds ratio*, ne sont pas la panacée. Car l'échelle logistique, sur laquelle se fonde la mesure par les *odds ratio*, est bien mieux adaptée aux cas où l'une des deux catégories contient beaucoup plus d'individus que l'autre. Un exemple : x_1 étant une variable binaire, supposons que la probabilité d'appartenance à C_1 soit de 4% lorsque $x_1 = 0$ et de 6% lorsque $x_1 = 1$. L'impact apparent de x_1 sur l'appartenance à C_1 est donc de 2 points. Ce gain de 2 points est, *dans l'absolu*, (très) faible. En revanche, il correspond à une augmentation *relative* de 50% de la probabilité, ce qui donne l'impression que l'impact est très important. Les *odds*

ratio combinent justement les deux aspects, une évolution faible dans l'absolu mais relativement importante [32]. Ce cas de figure est fréquent en épidémiologie lorsqu'elle analyse des maladies (relativement) rares, ce qui explique pourquoi les *odds ratio* y soient bien adaptés.

I.6.d Significativité statistique des effets marginaux

Il reste à calculer les écarts-types des grandeurs (31) ou (28) pour en apprécier la précision en calculant des intervalles de confiance (voir section I.3.b). Pour ce faire, on utilise une méthode, appelée *méthode delta*, traduction littérale de l'appellation anglo-saxonne *delta method*. Le principe est le suivant.

On a estimé un jeu de paramètres β_k ($k = 0, 1, \ldots, K$). On connaît la matrice de variance covariance des estimateurs, notée $Var(\hat{\beta})$. On s'intéresse à la grandeur θ, qui est une fonction connue des paramètres β_k : $\theta = h(\beta)$. Une valeur estimée de θ est donnée par : $\hat{\theta} = h(\hat{\beta})$. Pour en obtenir la variance, on applique la formule :

$$\hat{\sigma}^2_{\hat{\theta}} = \frac{\partial h(\hat{\beta})}{\partial \hat{\beta}'} \, Var(\hat{\beta}) \, \frac{\partial h(\hat{\beta})}{\partial \hat{\beta}} \tag{32}$$

où $\partial h(\hat{\beta})/\partial \hat{\beta}'$ (*resp.* $\partial h(\hat{\beta})/\partial \hat{\beta}$) est le vecteur ligne (*resp.* colonne) des dérivées partielles de h par rapport aux β_k.

Un exemple d'application, pour illustrer la méthode. Supposons que notre grandeur d'intérêt soit $\theta = \beta_2/\beta_1$, où β_1 et β_2 sont deux paramètres du modèle. Leurs valeurs estimées sont $\hat{\beta}_1 = 0.5$ et $\hat{\beta}_2 = 0.75$. Leurs écarts-types sont respectivement 0.2 et 0.3 (les variances respectives sont donc 0.04 et 0.09), et la covariance de $\hat{\beta}_1$ et $\hat{\beta}_2$ est de 0.01. La matrice de variance covariance est donc égale à :

$$Var(\hat{\beta}_1, \hat{\beta}_2) = \begin{pmatrix} 0.04 & 0.01 \\ 0.01 & 0.09 \end{pmatrix}$$

Une valeur estimée de θ est : $\hat{\theta} = \hat{\beta}_2/\hat{\beta}_1 = 0.75/0.5 = 1.5$. Les dérivées partielles de $h(\hat{\beta}_1, \hat{\beta}_2)$ sont égales à :

$$\frac{\partial h(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = -\frac{\hat{\beta}_2}{\hat{\beta}_1^2} = -\frac{0.75}{0.25} = -3 \quad \text{et} \quad \frac{\partial h(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.5} = 2$$

Par conséquent, le carré de l'écart-type de $\hat{\theta}$, donné par (32), est égal à :

$$\hat{\sigma}^2_{\hat{\theta}} = \begin{pmatrix} -3 & 2 \end{pmatrix} \begin{pmatrix} 0.04 & 0.01 \\ 0.01 & 0.09 \end{pmatrix} \begin{pmatrix} -3 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 & 2 \end{pmatrix} \begin{pmatrix} -0.10 \\ 0.15 \end{pmatrix} = 0.60$$

Il s'ensuit que l'écart-type de $\hat{\theta}$ est à peu près égal à 0.775.

Pour calculer les écarts-types des effets marginaux connaissant la matrice de va-

32. Le contraste logistique – voir section I.1.b – traite aussi cet aspect des choses.

riance covariance des $\hat{\beta}_k$, il faut calculer leurs dérivées partielles par rapport aux $\hat{\beta}_k$ puis appliquer la formule (32). Les calculs ne sont pas reproduits ici, mais ils sont intégrés dans la macro SAS utilisée dans la partie suivante du document.

II. Le modèle Logit : application

II.1 Introduction : remarques générales

La suite du document est consacrée au traitement complet d'un exemple. Il s'agit de mettre en pratique ce qu'a détaillé la première partie : estimer les paramètres du modèle, juger sa qualité, réaliser des tests d'hypothèse, ... On accordera une attention particulière à la présentation des résultats, surtout s'ils sont destinés à un public dépassant largement le périmètre des connaisseurs de l'outil.

Avant de présenter l'exemple d'application, on insistera sur deux points dont l'importance est souvent sous-estimée. Le premier concerne la spécification du modèle, c'est-à-dire le choix – raisonné – des variables. Le second a trait à l'utilisation d'une expression – *toutes choses égales par ailleurs* – censée caractériser le travail empirique effectué avec un modèle de type logit, qui peut induire le lecteur en erreur.

II.1.a Choix et organisation des variables

Le choix des variables est étroitement lié à la question que l'on se pose et qui motive l'utilisation du modèle logit. Cela concerne avant tout, bien entendu, les variables x, mais aussi la variable y, comme le montre le cas suivant.

Lorsqu'on étudie les différences de salaires entre les hommes et les femmes, on est souvent amené à spécifier un modèle de régression linéaire qui explique le niveau de la rémunération par la variable de sexe et plusieurs autres – le niveau de diplôme, la quotité de travail, le secteur d'activité, ... – ayant le statut de variables de contrôle. Les estimations du modèle donnent systématiquement une relation négative entre le fait d'être une femme et le niveau de salaire. Supposons maintenant qu'on s'intéresse à la relation inverse, et qu'on cherche à « expliquer » par un modèle logit le sexe de la personne (variable binaire) par son niveau de salaire. Cela paraît *a priori* très étrange voire absurde. C'est pourtant tout à fait justifié si on se place dans un contexte particulier, celui de l'*imputation de valeurs manquantes*. Supposons en effet qu'on ait collecté un ensemble d'informations sur un échantillon de salariés. On s'aperçoit que la variable de sexe n'est pas toujours renseignée. On souhaite pourtant faire des analyses sur l'ensemble de l'échantillon, sans éliminer les observations pour lesquelles on ne sait pas si l'individu est un homme ou une femme. Une solution est de commencer par estimer un modèle logit « expliquant » le sexe par le maximum de variables qui lui sont *a priori* corrélées. L'estimation se fait sur les seuls individus de l'échantillon pour lesquels on dispose de toutes les informations. Une fois connues les valeurs des paramètres du modèle, on est capable de prédire, pour chaque individu de sexe inconnu de l'échantillon, la probabilité qu'il soit un homme ou une femme et d'imputer une valeur (par exemple 1 pour un homme et 2 pour une femme) à la variable de sexe.

Mis à part ce cas très particulier, la question du choix se pose avant tout pour les x : quelles variables doit-on introduire dans le modèle et pourquoi ? Cela dépend de la finalité de la modélisation, et en particulier de la hiérarchie que l'on établit au sein des variables x.

Si on s'intéresse à une caractéristique individuelle particulière et à son rôle dans

l'appartenance à C_1 ou C_0, alors le modèle sera construit « autour » de cette variable particulière, appelée ici *variable principale*. Il faudra la distinguer des autres caractéristiques introduites dans le modèle, qui auront le statut de variables de contrôle car leur fonction sera avant tout de contrôler ces *effets de structure* (ou *effets de composition*) dont on a parlé à plusieurs reprises. Dans l'exemple introductif en avant-propos du document, la variable de nationalité est la variable principale, et le modèle logit est utilisé pour contrôler les *effets de structure* dus, entre autres, aux différences de niveaux de diplôme entre Français et étrangers, le diplôme étant alors considéré comme une variable de contrôle. La question à laquelle on cherchera à répondre est : quel rôle joue la variable principale dans l'appartenance des individus aux catégories C_0 et C_1, compte tenu du fait que ces deux sous-populations ne se ressemblent pas ?

S'il s'agit de mettre en évidence les traits distinctifs des deux catégories C_0 et C_1, c'est-à-dire de mener une analyse discriminante (voir section I.1.d), alors on doit traiter toutes les variables x au même niveau et ne pas instaurer de hiérarchie *a priori* entre elles. Le modèle sera utilisé pour répondre à une question du type : sur quelle(s) caractéristique(s) se distinguent fondamentalement les deux catégories d'individus ?

Indépendamment du mode d'utilisation du logit – analyse discriminante sans priviligier *a priori* de variable, ou bien analyse centrée sur une variable principale – il est rare qu'on parvienne du premier coup à trouver les bonnes caractéristiques x à introduire dans le modèle. Il est parfois – si ce n'est souvent – nécessaire de faire des ajustements en fonction de ce que produit l'estimation du modèle. Le choix des variables doit aussi être guidé par la facilité à présenter les résultats, qui conditionne largement leur lisibilité.

II.1.b *Toutes choses égales par ailleurs*, une expression à éviter

Une des difficultés de l'usage du modèle logit est de résister à ce qu'on appelle la « tentation de la causalité ». On entend par là l'utilisation de l'outil dans sa dimension « explicative » – voir la section I.1.c – dans l'objectif d'estimer l'« effet pur » de telle ou telle variable (x_1, x_2, ...) sur la variable catégorielle y. Ceci est particulièrement malvenu lorsque la variable principale, x_1 mettons, est une variable décrivant ou mesurant un comportement et peut être suspectée d'endogénéité. L'endogénéité se produit, rappelons-le, lorsque des facteurs inobservés, non pris en compte dans le modèle et qui ne figurent donc pas dans la liste des variables x, sont susceptibles d'influencer à la fois x_1 et y. On a cité – voir la section I.1.d – l'exemple de la variable caractérisant le secteur d'enseignement (public ou privé) d'un établissement dans un modèle cherchant à expliquer son impact dans la réussite scolaire. On le redit, ignorer ces phénomènes d'endogénéité peut conduire à des conclusions erronées, voire dans le pire des cas contraires à la réalité.

L'expression « toutes choses égales par ailleurs » est une autre manière d'exprimer cette idée de causalité, surtout si on la prend au pied de la lettre, c'est-à-dire si on

considère que dans les « choses » en question il y a aussi bien des caractéristiques observées et figurant dans la liste des variables x, que des caractéristiques inobservées ou inobservables. Là aussi, il convient d'être très prudent dans l'usage de l'expression. Car il ne faut pas oublier que les résultats des estimations sont *conditionnels* à la liste des variables x introduites dans le modèle, c'est-à-dire qu'ils dépendent des variables introduites. Ils peuvent varier, parfois substantiellement, si on en ajoute ou si on en retire. L'exemple d'application traité dans les pages suivantes va l'illustrer parfaitement.

II.1.c Présentation de l'exemple d'application

L'exemple retenu ici pour appliquer le modèle logit s'appuie sur une exploitation du panel 1995 de la Direction de l'Évaluation, de la Prospective et de la Performance (DEPP) du ministère en charge de l'éducation nationale, panel qui suit sur longue période une cohorte d'environ 18 000 élèves entrés en 6ème en 1995. L'échantillon d'étude contient 13 500 élèves de France métropolitaine qu'on a pu suivre jusqu'en classe de seconde, qui sont passés par une troisième générale et dont les variables utilisées dans l'analyse (cf *infra*) ont été correctement renseignées. C'est le logiciel SAS avec sa procédure `logistic` qui est utilisé.

Dans un premier temps, on souhaite étudier le rôle joué par l'éducation prioritaire dans l'orientation des élèves en fin de troisième. Plus précisément, on se demande si le fait pour un élève de troisième générale d'être dans un établissement situé en zone d'éducation prioritaire est un avantage ou au contraire – et comme on le pense souvent – un handicap pour passer en seconde générale (ou technologique). Pour cette étude, la variable d'intérêt est la variable binaire qui distingue les élèves qui ont été affectés en seconde générale (catégorie C_1) et ceux qui l'ont été en seconde professionnelle (catégorie C_0). La variable principale est la variable binaire qui distingue les éléves en éducation prioritaire et les autres.

On va le voir, il y a des *effets de structure* qu'il faut démêler pour répondre correctement à la question posée. Le recours à un modèle *logit* est de ce fait justifié. Pour contrôler les *effets de structure*, les variables (de contrôle) on été regroupées en quatre types : des caractéristiques démographiques de l'élève (année de naissance et sexe), son niveau à l'entrée en sixième, son milieu social et l'académie où il étudie.

Après avoir présenté quelques statistiques descriptives bien choisies de manière à justifier le recours au *logit*, on se livre ensuite à plusieurs estimations en introduisant une à une les variables de contrôle. Cette démarche est à visée pédagogique et n'a pas à être adoptée lors d'une étude. Il s'agit ici de bien comprendre les rôles respectifs des différentes variables de contrôle et de faire prendre conscience de l'ambiguïté de l'expression *toutes choses égales par ailleurs*.

Dans un second temps, on ne privilégiera pas de variable. La variable d'éducation prioritaire sera traitée au même niveau que les autres. On utilisera le modèle comme un outil d'analyse discriminante pour répondre à une question du type : parmi toutes les variables à notre disposition, quelle est celle ou quelles sont celles qui joue(nt)

le(s) premier(s) rôle(s) dans l'orientation en fin de 3ème ?

II.2 Premières statistiques descriptives

Les données sont conservées dans une table SAS, appelée ici `tab`. La variable d'intérêt, nommée `secondeg`, est la variable binaire distinguant les élèves qui ont été orientés en seconde générale (`secondeg=1`) à l'issue de leur troisième, et ceux qui ont suivi la voie professionnelle (`secondeg=0`). La variable principale est la variable binaire `zep` qui vaut 1 si l'élève étudie en zone d'éducation prioritaire (11,4% des élèves), et 0 sinon.

Pour poser le problème, on commence par croiser la variable d'intérêt et la variable principale, en utilisant la procédure `freq` de SAS :

```
proc freq data=tab;
  table zep*secondeg;
run;
```

La table 1 qui s'en déduit donne la part des élèves de troisième qui passent en seconde générale selon qu'ils étudient ou non en zone d'éducation prioritaire. On constate un écart de 13,5 points, dans le taux de passage en seconde générale, entre les élèves en ZEP et ceux hors ZEP, au bénéfice de ces derniers. Apparemment, étudier en ZEP diminuerait les chances de passer en seconde générale.

Table 1. Taux de passage en seconde générale selon la zone de l'établissement

	Part des élèves orientés en seconde générale (%)
ZEP	55,3
Hors ZEP	68,8
Ensemble	67,3

Lecture : 55,3% des élèves de troisième étudiant en ZEP sont passés en seconde générale.
Source : DEPP – Panel 1995.

Mais les élèves en ZEP et ceux hors ZEP ne se ressemblent pas. Pour le voir, il suffit de croiser la variable `zep` avec la variable `retard`, qui vaut 1 si l'élève a au moins un an de retard à l'entrée en 6ème, et 0 sinon :

```
proc freq data=tab;
  table zep*retard;
run;
```

La table 2 montre que les élèves de ZEP sont, en proportion, deux fois plus nombreux à être entrés en sixième avec au moins un an de retard.

Table 2. Retard en sixième selon la zone de l'établissement

	Part des élèves en retard en sixième (%)
ZEP	22,8
Hors ZEP	11,3
Ensemble	12,6

Lecture : 22,8% des élèves de troisième étudiant en ZEP sont entrés en sixième avec au moins un an de retard.
Source : DEPP – Panel 1995.

Or, d'une manière générale, que l'on étudie ou non en ZEP, être entré en retard en sixième diminue sensiblement les chances de se retrouver en seconde générale. On le constate en croisant les variables `secondeg` et `retard` :

```
proc freq data=tab;
  table retard*secondeg;
run;
```

Un quart seulement des entrants en sixième avec retard passe en voie générale, contre quasiment les trois-quarts des élèves à l'heure ou en avance (table 3).

Table 3. Taux de passage en seconde générale selon le retard en sixième

	Part des élèves orientés en seconde générale (%)
En retard en 6ème	25,3
A l'heure ou en avance en 6ème	73,3
Ensemble	67,3

Lecture : 25,3% des élèves entrés en retard en 6ème sont passés en seconde générale.
Source : DEPP – Panel 1995.

En conséquence, à partir du moment où les élèves en ZEP sont plus souvent en retard que les autres et où le retard scolaire est un désavantage dans l'orientation post-troisième, il n'est pas étonnant de constater que les élèves de ZEP passent moins souvent que les autres en seconde générale. Au moins une partie des 13,5 points

d'écart (table 2) s'explique ainsi par la différence de composition des populations d'élèves de ZEP d'une part, hors ZEP d'autre part.

Pour le voir, on dédouble la table 1, en isolant les élèves ayant au moins un an de retard d'un côté, et les élèves à l'heure ou en avance de l'autre :

```
proc freq data=tab(where=(retard=1));
  table zep*secondeg;
run;
proc freq data=tab(where=(retard=0));
  table zep*secondeg;
run;
```

La table 4 montre que l'écart, dans le taux de passage, entre les élèves de ZEP et les autres s'établit à 1,3 point pour les élèves en retard et à quasiment 10 points pour ceux à l'heure ou en avance. Par conséquent, le retard en sixième explique une partie – mais seulement une partie – de l'écart constaté dans la table 1. Compte tenu de l'âge à l'entrée en sixième, l'écart n'est plus de 13,5 points mais d'un pourcentage compris entre 1,3 point et 9,8 points.

Table 4. Taux de passage en seconde générale selon la zone de l'établissement et le retard en sixième

	Élèves en retard	Élèves à l'heure ou en avance
ZEP	24,2	64,5
Hors ZEP	25,5	74,3
Ensemble	25,3	73,3

Lecture : 24,2% des élèves de troisième étudiant en ZEP et qui sont entrés en retard en sixième sont passés en seconde générale.
Source : DEPP – Panel 1995.

D'autres facteurs jouent, comme la catégorie sociale de l'élève. Ainsi, 4% des élèves de troisième en établissement ZEP sont des filles ou fils de cadres, alors que dans les établissements ne relevant pas de l'éducation prioritaire, la proportion est de 18%. Or, 91% des enfants de cadres passent en seconde générale contre 63% pour les enfants d'autres milieux sociaux. En conséquence, si les élèves de ZEP sont moins fréquemment orientés en voie générale, c'est en partie parce qu'ils vivent plus souvent dans des milieux socialement défavorisés, qu'ils sont moins souvent portés par des familles ayant les ressources pour les aider.

On tient donc là une autre explication possible de l'écart des taux de passage ZEP/hors ZEP de la table 1 : il serait aussi dû à la différence de structure sociale des deux populations ZEP et hors ZEP.

Pour la neutraliser, on doit distinguer les élèves non seulement selon leur retard à l'entrée en sixième mais aussi selon leur milieu social. On est donc amené à éditer la table 1 pour quatre sous-populations (en retard/à l'heure croisé avec cadre/non cadre), donc à créer quatre tables, dont il faut faire la synthèse pour répondre à la question du lien entre zone d'éducation et passage en seconde générale. Avec un critère binaire supplémentaire distinguant deux grandes catégories d'élèves, cela ferait 8 tables. Et ainsi de suite.

On voit que cette manière de faire est impraticable. Elle l'est encore plus si on introduit des critères à plusieurs modalités, si, par exemple, on caractérise le milieu social de l'élève en distinguant plus finement les catégories sociales au lieu de s'en tenir à la dichotomie cadre/non cadre. Et si on ajoute des variables continues, comme le niveau de l'élève en sixième, cela devient pratiquement infaisable.

Il est donc nécessaire de se tourner vers un outil comme le modèle logit, qui permet ici de savoir deux choses : à milieu social et âge d'entrée en sixième donnés, un élève de ZEP a-t-il toujours moins de chances qu'un autre de passer en seconde générale ? Et si oui, à combien se chiffre son handicap ? La table 1 l'évalue à 13,5 points, mais sans tenir compte des spécificités des élèves en ZEP en termes d'âge d'entrée en sixième et de milieu social. Que devient cette différence si on les prend en compte ?

II.3 Spécifications du modèle et estimation

Tout au long de cette section II.3, on va progressivement enrichir le modèle en introduisant les variables de contrôle les unes après les autres.

II.3.a Introduction de la variable d'âge à l'entrée en sixième

On commence par la variable donnant l'âge à l'entrée en sixième. On dispose, dans la source de données, de la variable `annais` qui nous indique l'année de naissance de l'élève. On choisit de la transformer en une variable distinguant trois catégories d'élèves : ceux nés avant 1984, ceux nés en 1984, ceux nés en 1985 ou après. Les premiers sont en retard d'au moins un an à l'entrée en sixième, les deuxièmes sont à l'heure et les troisièmes en avance.

Dans une étape *data* de SAS, on crée donc trois variables binaires à partir de la variable `annais` :

```
retard=(annais<1984);
alheure=(annais=1984);
avance=(annais>1984);
```

On a affaire ici à une variable polytomique ordonnée à trois modalités (voir section I.2), dont il faut choisir une modalité qui sera considérée comme référence. Conformément à ce qu'on a préconisé page 23, on pourrait retenir la première (i.e. les élèves en retard). Pour faciliter les commentaires, on prendra plutôt la modalité qui correspond à la « norme », c'est-à-dire les élèves nés en 1984 (variable `alheure`).

Pour estimer le modèle *logit* avec la variable d'éducation prioritaire et la variable d'âge à l'entrée en sixième, on écrit les instructions suivantes :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance ;
run;
```

L'option `descending` est indispensable. Elle assure que les paramètres estimés sont bien ceux du modèle pour lequel la valeur 1 de la variable `secondeg` correspond au passage en seconde générale[33]. La variable binaire `zep` est introduite telle quelle. Pour la variable d'âge à l'entrée en sixième, on n'introduit pas l'indicatrice représentant la modalité de référence (voir section I.2.c).

L'exécution de la procédure produit les résultats reportés ci-dessous. La partie intéressante est intitulée `Analysis of Maximum Likelihood Estimates`. Elle présente les résultats des estimations. La colonne `Parameter` donne le nom des variables introduites dans le modèle (`Intercept` est le nom du terme constant), la colonne `Estimate` donne les valeurs estimées des paramètres associés aux variables du modèle, la colonne `Standard Error` en donne les écarts-types, et la colonne `Pr > Chisq`

33. Il s'agit là d'une bizarrerie de SAS : au lieu de coder la variable catégorielle en 1/0 comme il est usuel de le faire, SAS la code par défaut en 1/2.

le seuil de significativité (voir section I.5), la colonne précédente présentant les valeurs de la statistique de test utilisée par défaut par SAS.

The LOGISTIC Procedure

Model Information

Data Set	WORK.TAB
Response Variable	secondeg
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	13499
Number of Observations Used	13499

Response Profile

Ordered Value	secondeg	Total Frequency
1	1	9081
2	0	4418

Probability modeled is secondeg=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	17071.165	15417.143
SC	17078.676	15447.185
-2 Log L	17069.165	15409.143

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1660.0223	3	<.0001
Score	1703.1205	3	<.0001
Wald	1324.0456	3	<.0001

```
                Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter    DF    Estimate         Error    Chi-Square    Pr > ChiSq

Intercept     1      1.0048        0.0222     2047.5954        <.0001
zep           1     -0.3831        0.0597       41.1646        <.0001
retard        1     -2.0170        0.0599     1134.1102        <.0001
avance        1      1.8369        0.1924       91.1643        <.0001


                        Odds Ratio Estimates

                     Point          95% Wald
                Effect    Estimate      Confidence Limits

                zep          0.682       0.606       0.766
                retard       0.133       0.118       0.150
                avance       6.277       4.305       9.152
```

La table 5 est la manière standard de reporter les résultats des estimations[34]. Il est ainsi usuel de positionner la significativité de chaque paramètre par rapport à trois seuils prédéfinis – en l'espèce, 1%, 5% et 10% – et de la représenter par des astérisques. Ainsi, l'adjonction de trois astérisques à la valeur du paramètre signifie qu'il est significatif au seuil de 1%. Avec deux astérisques, il n'est pas significatif au seuil de 1%, mais l'est au seuil de 5%. Lorsqu'il n'y a qu'un astérisque, le paramètre est significatif au seuil de 10% mais pas au seuil de 5%. Enfin, en l'absence d'astérisque, il n'est pas significatif au seuil de 10%. Dernière précision, la modalité de référence de chaque variable qualitative (binaire ou polytomique) est rappelée par la mention (*ref*= ...) attachée au libellé de la variable.

Table 5. Ajout de la variable *Age à l'entrée en sixième*

Variable	Paramètre estimé	Écart-type
Constante	1,005***	0,022
Appartenance à une ZEP (*ref=non*)		
oui	−0,383***	0,060
Âge à l'entrée en sixième (*ref=à l'heure*)		
en retard	−2,017***	0,060
en avance	1,837***	0,192

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

Le signe d'un paramètre associé à une variable indique dans quel sens influe cette variable sur la variable d'intérêt. Ainsi, le paramètre de la variable `zep` est négatif :

34. Nous verrons en section II.4 comment présenter les résultats à un public moins initié.

être en éducation prioritaire influe négativement sur le passage en seconde générale. Les deux autres indicatrices, représentant les modalités *en retard* et *en avance* de la variable d'âge à l'entrée en sixième, s'interprétent en regard de la modalité de référence (*être à l'heure*) : être en retard est, par rapport au fait d'être à l'heure, pénalisant pour passer en seconde générale (le paramètre de l'indicatrice `retard` est négatif) ; en revanche, être en avance est un avantage, toujours par rapport au fait d'être à l'heure (le paramètre de l'indicatrice `avance` est positif).

Tous les paramètres sont significatifs au seuil de 1%. Cela signifie qu'on a moins de 1% de risques de se tromper en affirmant que ces paramètres sont différents de 0 (voir section I.5.a). En réalité, si on se reporte à la sortie SAS (page 60 et suivante), le risque de se tromper est beaucoup plus faible : moins de 1/10000 (voir la colonne `Pr > Chisq`). On peut donc affirmer sans crainte que les trois indicatrices influent sur l'orientation post-troisième.

En conclusion, la zone d'éducation prioritaire joue négativement sur l'orientation en seconde générale même en tenant compte de l'âge d'entrée en sixième.

II.3.b Ajout de la distinction fille/garçon

On poursuit en introduisant la variable renseignant le sexe de l'élève. On aura préalablement créé la variable `fille`, qui vaut 1 si l'élève est une fille et 0 sinon, et qui comme toute variable binaire est introduite en l'état dans l'instruction `model` de la procédure. On relance donc la procédure *logistic* :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance fille;
run;
```

Les résultats, issus de la sortie SAS, sont mis en forme et reportés dans la table 6.

Table 6. Ajout de la variable *Sexe de l'élève*

Variable	Paramètre estimé	Écart-type
Constante	0,806***	0,029
Appartenance à une ZEP (*ref=non*)		
oui	−0,382***	0,060
Age à l'entrée en sixième (*ref=à l'heure*)		
en retard	−2,035***	0,060
en avance	1,845***	0,193
Sexe de l'élève (*ref=garçon*)		
fille	0,400***	0,040

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

Les filles, à zone d'éducation et âge en sixième donnés, vont davantage en seconde générale que les garçons. On remarque que les valeurs des paramètres des variables

déjà présentes dans le modèle ne changent quasiment pas (sauf celle de la constante). C'est signe que la variable *Sexe de l'élève* n'est pas liée aux variables de zone d'éducation prioritaire et d'âge à l'entrée en sixième. Il y a, à peu de choses près, autant de filles que de garçons en zone d'éducation prioritaire, et l'âge en sixième des filles et des garçons est le même ou peu s'en faut. On le vérifie, sur les données, en croisant ces variables entre elles.

II.3.c Ajout du milieu social de l'élève

La source de données contient la variable `pcschef` à 7 modalités, numérotées de 1 à 7, qui repère la catégorie sociale du chef de ménage [35]. La modalité 7 regroupe les cas où le chef de famille a déclaré être sans activité et les quelques autres où il n'a pas répondu à la question sur sa catégorie sociale d'appartenance, où par conséquent la variable de milieu social est « à valeurs manquantes ».

Avant de poursuivre, il convient de s'arrêter un moment sur le traitement général des variables à valeurs manquantes, c'est-à-dire celles qui ne sont pas renseignées alors qu'elles devraient l'être.

Une solution serait d'éliminer les observations où le cas se présente. Ainsi, on aurait pu supprimer de notre analyse tous les élèves dont on ignore la catégorie sociale de ses parents. C'est une solution qui doit être évitée, pour plusieurs raisons. D'abord, elle fait diminuer la taille de l'échantillon d'étude, avec la perte de précision des estimations que cela implique (voir page 28). Ensuite, si les cas de valeurs manquantes ne sont pas distribués au hasard, supprimer les observations concernées risquerait de conduire à un échantillon qui ne serait plus représentatif de la population étudiée et qu'il faudrait alors redresser. Enfin, lorsque c'est une variable de contrôle [36] qui est concernée, il faut rappeler qu'elle est somme toute secondaire par rapport à la variable principale, et que c'est le rôle joué par la variable principale qui importe avant tout.

Plusieurs solutions alternatives sont envisageables. Un premier traitement possible consiste à créer une modalité « non renseignée » et la variable indicatrice qui va avec, à condition que le nombre de valeurs manquantes soit suffisamment élevé. Une autre possibilité est de regrouper les valeurs manquantes avec une autre modalité de la variable que l'on pense *a priori* être proche. C'est ce qui a été fait avec les absences de réponse à la question du milieu social, qui ont été absorbées dans la catégorie 7 (sans activité). Des analyses sur d'autres données conduites par ailleurs ont en effet montré que ces non répondants ont un profil très proche des inactifs.

Il y a en pratique deux manières d'intégrer la variable polytomique `pcschef` à l'analyse. Commençons par la plus économe en programmation. Elle consiste à faire traiter par la procédure `logistic` la transformation de la variable `pcschef` en in-

35. On verra – section II.9 – que l'ajout de la catégorie sociale comme variable de contrôle ne va pas de soi.

36. S'il manque des valeurs à la variable principale, alors il faut se résoudre à supprimer les observations correspondantes, avec tous les inconvénients que cette suppression implique.

dicatrices. Pour ce faire, on ajoute l'instruction `class` en précisant la modalité de référence – on a pris ici la modalité 7 « *sans activité professionnelle et non répondants* » – de la manière suivante (Attention ! La variable mise dans l'instruction `class` doit obligatoirement être en format caractère) :

```
proc logistic data=tab descending ;
  class pcschef (ref='7') / param=ref;
  model secondeg = zep retard avance fille pcschef;
run;
```

On obtient en sortie :

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1771	0.1260	1.9766	0.1597
zep	1	-0.0627	0.0622	1.0161	0.3134
retard	1	-1.8560	0.0623	888.0160	<.0001
avance	1	1.6013	0.1957	66.9244	<.0001
fille	1	0.4959	0.0412	145.0180	<.0001
pcschef 1	1	0.7561	0.1664	20.6380	<.0001
pcschef 2	1	0.9170	0.1389	43.5526	<.0001
pcschef 3	1	2.3521	0.1452	262.2744	<.0001
pcschef 4	1	1.4266	0.1330	115.1187	<.0001
pcschef 5	1	0.6788	0.1311	26.8223	<.0001
pcschef 6	1	0.3000	0.1269	5.5872	0.0181

Cette manière de procéder permet d'obtenir directement, sans instruction supplémentaire, le résultat du test de nullité jointe des paramètres associés aux modalités de la variable de milieu social (voir section I.5.b, page 37). Il se trouve dans la partie de la sortie standard de SAS intitulée `Type 3 Analysis of Effects`. La dernière ligne donne la valeur de la statistique de test et le seuil de significativité :

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
zep	1	1.0161	0.3134
retard	1	888.0160	<.0001
avance	1	66.9244	<.0001
fille	1	145.0180	<.0001
pcschef	6	833.1727	<.0001

Le seuil de significativité (`Pr > ChiSq`) nous dit que le risque de nous tromper en affirmant que les 6 paramètres ne sont pas tous égaux à 0 est inférieur à 1/10000. On peut donc affirmer que le milieu social joue (globalement) un rôle dans l'orientation en seconde.

La seconde méthode pour estimer les paramètres associés aux modalités de la variable de milieu social, conforme à la démarche générale (voir section I.2.c), est de créer 7 indicatrices et d'en introduire 6 dans le modèle, en excluant celle représentant la modalité de référence. Les 7 indicatrices, nommées `csp1` à `csp7`, sont obtenues par les instructions suivantes à placer dans une étape *data* :

```
array csp(i) csp1-csp7;
do i=1 to 7;csp=(i=pcschef*1);end;
```

Par exemple, la variable `csp5` vaut 1 si l'élève est fille ou fils d'employé (modalité 5 de la variable `pcschef`), et 0 sinon. L'estimation des paramètres se fait par les instructions suivantes (la modalité de référence `csp7` est exclue de l'instruction `model`) :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance fille csp1-csp6;
run;
```

On obtient en sortie :

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1771	0.1260	1.9766	0.1597
zep	1	-0.0627	0.0622	1.0161	0.3134
retard	1	-1.8560	0.0623	888.0160	<.0001
avance	1	1.6013	0.1957	66.9244	<.0001
fille	1	0.4959	0.0412	145.0180	<.0001
csp1	1	0.7561	0.1664	20.6380	<.0001
csp2	1	0.9170	0.1389	43.5526	<.0001
csp3	1	2.3521	0.1452	262.2744	<.0001
csp4	1	1.4266	0.1330	115.1187	<.0001
csp5	1	0.6788	0.1311	26.8223	<.0001
csp6	1	0.3000	0.1269	5.5872	0.0181

On pourrait prendre la modalité 6 (« *ouvriers* ») comme modalité de référence, au motif que c'est la plus fréquente. Dans ce cas, on écrit :

```
proc logistic data=tab descending ;
  class pcschef (ref='6') / param=ref;
  model secondeg = zep retard avance fille pcschef;
run;
```

ou bien :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance fille csp1-csp5 csp7;
run;
```

qui produit :

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.1229	0.0412	8.9102	0.0028
zep	1	-0.0627	0.0622	1.0161	0.3134
retard	1	-1.8560	0.0623	888.0160	<.0001
avance	1	1.6013	0.1957	66.9244	<.0001
fille	1	0.4959	0.0412	145.0180	<.0001
csp1	1	0.4561	0.1166	15.2926	<.0001
csp2	1	0.6170	0.0724	72.5420	<.0001
csp3	1	2.0521	0.0837	601.4371	<.0001
csp4	1	1.1266	0.0602	350.4400	<.0001
csp5	1	0.3788	0.0561	45.5295	<.0001
csp7	1	-0.3000	0.1269	5.5872	0.0181

Avec cette autre modalité de référence, seules les valeurs des paramètres associés aux catégories sociales, ainsi que celle du paramètre du terme constant (`Intercept`), sont modifiées. On vérifie toutefois que les écarts entre les différentes modalités de la variable `pcschef` ne changent pas. Par exemple, la différence entre les employés (modalité 5) et les ouvriers (modalité 6) est de 0,6788-0,3000=0,3788 dans le premier cas (i.e. modalité de référence 7) et de 0,3788-0=0,3788 dans le second cas.

L'avantage de la méthode consistant à créer explicitement les indicatrices est qu'elle permet de faire des tests autres que celui de la nullité jointe des paramètres associés à la variable `pcschef`. On peut notamment réaliser un test d'égalité de deux paramètres, par exemple ceux des catégories 1 (agriculteurs exploitants) et 2 (artisans, commerçants et chefs d'entreprise). On utilise pour ce faire l'instruction `test` de la procédure :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance fille csp1-csp6;
  test csp1=csp2;
run;
```

On obtient :

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
Test 1	1.5581	1	0.2119

Si on rejette l'hypothèse nulle de l'égalité des deux paramètres, on a plus de 20% de chances de se tromper. Il est donc préférable de ne pas le faire et d'affirmer que, sur notre échantillon, on ne distingue pas de différence dans l'orientation post-troisième entre les enfants d'agriculteurs et les enfants d'artisans, commerçants et

chefs d'entreprise pourvu qu'ils soient de même sexe, aient le même âge à l'entrée en sixième et soient dans le même secteur d'enseignement (éducation prioritaire ou non).

La table 7 met en forme les résultats. Le résultat le plus spectaculaire est la modification substantielle du paramètre de la variable de zone d'éducation. Il est toujours négatif mais n'est pas significatif au seuil de 10%. En se reportant à la sortie de la procédure ci-dessus, on constate même qu'il ne l'est pas au seuil de 30% (`Pr > ChiSq = 0.3134`). À sexe, âge en sixième et milieu social donnés, la zone d'étude ne semble[37] pas jouer de rôle dans l'affectation en seconde générale. La catégorie sociale des parents de l'élève est ainsi responsable d'*effets de structure* d'ampleur importante. On a déjà vu (page 57) que les enfants de cadres sont sous-représentés en ZEP et qu'ils poursuivent plus souvent leur scolarité en seconde générale. Ce tout dernier point est confirmé par la valeur du paramètre associé à la modalité *Cadres, Professions intellectuelles supérieures* (table 7), valeur élevée relativement à celles des paramètres des autres modalités.

Table 7. Ajout de la variable *Milieu social de l'élève*

Variable	Paramètre estimé	Écart-type
Constante	−0,177	0,126
Appartenance à une ZEP (*ref=non*)		
oui	−0,063	0,062
Age à l'entrée en sixième (*ref=à l'heure*)		
en retard	−1,856***	0,062
en avance	1,601***	0,196
Sexe de l'élève (*ref=garçon*)		
fille	0,496***	0,041
Milieu social de l'élève (*ref=sans act. prof., non rép.*)		
agriculteurs exploitants	0,756***	0,166
artis., commerç., chefs d'entrep.	0,917***	0,139
cadres, prof. intell. sup.	2,352***	0,145
prof. intermédiaires	1,427***	0,133
employés	0,679***	0,131
ouvriers	0,300**	0,127

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

Il y a un autre résultat, plutôt surprenant : la valeur du paramètre associé à la variable *Sexe de l'élève* a sensiblement changé après l'ajout du milieu social, indiquant qu'il y aurait un lien entre sexe et milieu social. Quand on regarde les choses d'un peu plus près et que l'on croise la variable de sexe avec la CSP, on s'aperçoit que les filles sont surreprésentées chez les employés et les ouvriers. Cela est

37. Restons prudents ! Voir page 35.

dû aux processus d'orientation qui ont lieu (ou avaient lieu à cette époque) au cours du premier cycle. Notamment, les élèves des quatrième et troisième technologiques se recrutent souvent parmi les fils d'employés ou d'ouvriers. Par conséquent, ceux-ci se retrouvent en moins grand nombre en troisième générale. Autrement dit, l'échantillon que nous avons sélectionné – les élèves qui sont passés par la troisième générale – n'est pas représentatif des entrants en sixième en 1995. Ceci peut produire ce qu'on appelle des *biais de sélection*, c'est-à-dire des résultats biaisés dus au fait que l'échantillon n'est pas représentatif, qu'il concerne une population qui a été sélectionnée. Il faut l'avoir en tête. Toutefois, la sélection n'est pas très marquée, en tout cas pas suffisamment pour remettre en cause les résultats présentés ici.

II.3.d Ajout du niveau de l'élève en 6ème

Deux variables, issues des épreuves nationales d'évaluation pour les élèves de 6ème, permettent d'avoir une idée du niveau de l'élève, en français et en mathématiques, à son entrée au collège. Il s'agit là de variables continues, nommées `fran` et `math`, qu'on ajoute en l'état à la liste des variables (voir section I.2) :

```
proc logistic data=tab descending ;
  model secondeg = zep retard avance fille csp1-csp6 fran math;
  test fran=math;
run;
```

On en profite pour faire un test d'égalité des paramètres associés aux deux variables de niveau (instruction `test`), test qui donne comme résultat :

```
                Linear Hypotheses Testing Results

                          Wald
Label       Chi-Square      DF    Pr > ChiSq

Test 1          0.4353       1        0.5094
```

Vu le seuil de significativité (`Pr > ChiSq`), on ne peut pas rejeter l'hypothèse d'égalité des paramètres : le niveau en français et celui en mathématiques ont la même force d'impact sur l'orientation [38].

La table 8 présente les résultats du modèle. L'ajout des variables de niveau de l'élève à son entrée en sixième change considérablement la donne. L'éducation prioritaire joue maintenant un rôle positif dans le passage en seconde générale [39]. Le paramètre associé à la variable `zep` est significatif au seuil de 1% (et même au seuil de 1/10000, d'après les sorties SAS non reproduites ici). Les *effets de structure*, en

38. ... conditionnellement aux autres variables du modèle ...

39. Ce type de résultat, mais appliqué à d'autres variables que le passage en seconde, a déjà été mis en évidence par Jean-Paul Caille : « Les collégiens de ZEP à la fin des années quatre-vingt-dix. Caractéristiques des élèves et impact de la scolarisation en ZEP sur la réussite », *Éducation et Formation*, n° 61, oct-déc 2001.

tout cas ceux captés par les variables introduites dans le modèle, sont tels que leur prise en compte inverse le signe du paramètre des premières estimations.

Table 8. Ajout des deux variables de niveau de l'élève

Variable	Paramètre estimé	Écart-type
Constante	−6,240***	0,202
Appartenance à une ZEP (*ref=non*)		
oui	0,489***	0,072
Age à l'entrée en sixième (*ref=à l'heure*)		
en retard	−1,278***	0,068
en avance	1,248***	0,208
Sexe de l'élève (*ref=garçon*)		
fille	0,521***	0,048
Milieu social de l'élève (*ref=sans act. prof., non rép.*)		
agriculteurs exploitants	0,370**	0,185
artis., commerç., chefs d'entrep.	0,642***	0,155
cadres, prof. intell. sup.	1,804***	0,161
prof. intermédiaires	1,009***	0,148
employés	0,445***	0,146
ouvriers	0,244*	0,142
Niveau en français (en sixième)	0,061***	0,003
Niveau en maths (en sixième)	0,064***	0,003

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

Autre remarque, l'ajout des deux variables de niveau modifie les valeurs des paramètres des autres variables, signe que les élèves ayant eu en sixième les meilleurs résultats ne vivent pas dans n'importe quelle famille.

II.3.e Ajout d'indicatrices académiques

Dernier enrichissement du modèle, l'introduction du niveau académique, c'est-à-dire de la variable – nommée `acad` – indiquant dans quelle académie l'élève a suivi sa scolarité de troisième. Comme pour le milieu social de l'élève (voir section II.3.c), il y a deux manières d'introduire la variable académique.

Si on retient la première – qui consiste à utiliser l'instruction `class` de la procédure `logistic` – en prenant comme académie de référence celle de Paris (`acad='01'`), on écrit :

```
proc logistic data=tab descending ;
  class pcschef (ref='7') acad (ref='01') / param=ref;
  model secondeg = zep retard avance fille pcschef fran math acad;
run;
```

Les résultats, non reproduits ici, ne modifient pas significativement les précédents.

On notera juste que la valeur du paramètre de la variable `zep` est un peu plus faible : $0,376$ au lieu de $0,489$ sans les indicatrices académiques.

Le résultat du test de nullité jointe des 25 paramètres associés aux académies nous dit que le risque de nous tromper en affirmant qu'ils ne sont pas tous égaux à 0 est inférieur à 1/10000. Le niveau académique joue bien un rôle dans l'orientation post-troisième.

L'inconvénient de ces variables indicatrices est qu'elles ne nous disent pas ce que l'on cherche à contrôler dans le modèle. Est-ce la politique académique d'orientation ? Ou bien le contexte économique ? Cela étant, ces indicatrices académiques sont ici des variables de contrôle, dont l'objectif premier est de contrôler l'hétérogénéité observée.

II.4 Calcul d'un effet marginal

Les tables des pages précédentes présentent les résultats de l'estimation sous la forme généralement utilisée. À ce stade, on sait dire si telle ou telle caractéristique joue un rôle positif ou négatif sur le passage en seconde générale : il suffit de lire le signe du paramètre concerné. On sait aussi mesurer notre degré de certitude lorsqu'on affirme que tel ou tel facteur compte en matière d'orientation en fin de troisième : on regarde le seuil de significativité (statistique) du paramètre. Par contre, la valeur du paramètre en tant que telle ne nous donne pas une idée immédiate de l'importance du facteur. Notamment, on ne sait pas mesurer l'influence de notre variable principale (l'éducation prioritaire) sur l'orientation en fin de 3ème. Il faut alors se tourner vers d'autres grandeurs statistiques, celles qui mesurent ce qu'on a appelé la significativité pratique des différents facteurs (section I.6).

L'*odds ratio* est la mesure la plus employée. Elle est automatiquement produite par la procédure. L'*odds ratio* figure à la fin de la sortie standard (voir page 60 et suivante), dans la partie `Odds Ratio Estimates`.

Avec le modèle complet sans les indicatrices académiques (i.e. celui de la section II.3.d), on obtient en sortie :

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
zep	1.631	1.416	1.878
retard	0.279	0.244	0.319
avance	3.483	2.319	5.232
fille	1.683	1.532	1.849
csp1	1.447	1.007	2.081
csp2	1.900	1.402	2.574
csp3	6.073	4.432	8.322
csp4	2.742	2.050	3.668
csp5	1.560	1.172	2.078
csp6	1.276	0.967	1.685
fran	1.063	1.056	1.070
math	1.066	1.061	1.072

L'*odds ratio* attaché à la variable `zep` est égal à 1,631, avec [1.416 , 1.878] comme intervalle de confiance à 95%. On l'a vu en section I.6.b, cela signifie précisément que la chance *relative* de passer en seconde générale est environ 1,6 fois plus élevée pour un enfant en ZEP que pour un enfant hors ZEP, conditionnellement aux facteurs pris en compte dans le modèle (i.e. à sexe, âge, milieu social et niveau en sixième fixés). Il est entendu que la chance relative est un rapport de probabilités : c'est la probabilité de passer en seconde générale rapportée à celle de ne pas y passer. L'*odds ratio* n'est donc pas un rapport de deux probabilités, mais un rapport de rapports de probabilités. Il ne faut surtout pas dire que le fait d'être en ZEP multiplie par 1,6 la probabilité de passer en seconde générale, à mêmes caractéristiques observées. On verra plus loin que ce résultat est, en ces termes, complètement faux.

La seconde solution est de calculer l'effet marginal de la variable `zep` (section I.6.c). Rappelons-en le principe :

- on « force » chaque élève de l'échantillon à étudier en ZEP : la valeur de la variable `zep` est mise systématiquement à 1 ; dans ce contexte, on calcule pour chaque élève la probabilité qu'il a d'être orienté en seconde générale ;
- on « force » chaque élève de l'échantillon à étudier hors ZEP : la valeur de la variable `zep` est mise systématiquement à 0 ; on calcule pour chaque élève la probabilité qu'il a d'être orienté en seconde générale ;
- on calcule, pour chaque élève, la différence entre ces deux probabilités ;
- on prend la moyenne, sur l'échantillon, de ces différences individuelles de probabilités.

L'effet marginal d'une variable qualitative, qu'elle soit binaire ou polytomique[40], se calcule grâce à la macro SAS *marginal*, détaillée en annexe du document. Elle compte quatre paramètres :

- `tab_ent` nomme la table SAS contenant les données individuelles, en entrée de la macro ;
- `x` donne la liste de toutes les variables introduites dans le modèle, *dans l'ordre où* elles l'ont été ;
- `param_ent` nomme la table SAS issue de l'exécution de la procédure *logistic*, qui contient les valeurs des paramètres estimés ainsi que la matrice de leurs variances et covariances ;
- `var_qual` nomme la variable (de nature qualitative) ou la liste des indicatrices qui lui sont associées, dont on veut calculer l'effet marginal ; s'il s'agit d'une variable binaire (comme la variable `zep`), alors la valeur du paramètre est le nom de la variable ; s'il s'agit d'une variable polytomique (comme l'âge à l'entrée en sixième), on met la liste des variables binaires représentant les modalités (sauf la modalité de référence) de la variable polytomique, *dans l'ordre où* elles apparaissent dans la liste `x`.

La macro calcule aussi l'écart-type de chaque effet marginal en utilisant la méthode delta (voir section I.6.d).

Pour calculer l'effet marginal de la variable `zep` avec la spécification du modèle de la section II.3.d, on procède comme suit. On définit d'abord, par une macro-variable appelée ici `listvar`, la liste des variables introduites dans le modèle :

```
%let listvar=zep retard avance fille csp1 csp2 csp3 csp4 csp5 csp6
             fran math ;
```

L'intérêt est de s'assurer ainsi que la liste des variables écrite dans la procédure *et* dans la macro (paramètre `x`) est strictement la même, condition nécessaire à sa bonne exécution. Attention ! Il ne faut pas utiliser la notation raccourcie `csp1-csp6`, sinon la macro ne s'exécute pas. Il faut écrire *toutes* les modalités de la variable de milieu social (exception faite de la modalité de référence).

40. On verra plus loin comment faire dans le cas d'une variable continue.

On lance ensuite la procédure `logistic` avec deux options supplémentaires : `outest=` et `covout`. La première permet de conserver dans une table SAS, que l'on nomme après le signe d'égalité, les valeurs des paramètres ainsi que leurs variances et covariances que l'on obtient grâce à la seconde option `covout`. On écrit donc :

```
proc logistic data=tab descending covout outest=param;
  model secondeg = &listvar;
run;
```

En faisant simplement appel à la macro variable `&listvar`, on introduit les variables dans le modèle selon l'ordre souhaité.

On est maintenant en mesure de calculer l'effet marginal de la variable `zep`, c'est-à-dire d'exécuter la macro qui le fait. Supposons que le fichier qui contient la macro se nomme `fichier1`. Supposons aussi qu'il soit conservé dans un répertoire nommé, mettons, `d:\Macro SAS`. On alloue d'abord le fichier contenant la macro à la *file* nommée ici `ff` :

```
filename ff 'd:\Macro SAS'; run;
```

Puis on fait appel à la macro par l'instruction :

```
%include ff(fichier1);
```

Enfin, on exécute la macro :

```
%marginal(tab_ent=tab,x=&listvar,
          param_ent=param,var_qual=zep);
```

En utilisant `&listvar`, on est certain d'avoir toutes la variables utilisées pour l'estimation et dans le même ordre. La table des paramètres estimés a le nom donné par l'option `outest=` de la procédure `logistic`. Enfin, `var_qual` désigne la variable pour laquelle on calcule l'effet marginal. On obtient en sortie :

Effet marginal de 'zep'			
	effet marginal	écart_type	significativité
ZEP	6.6174	0.9169	< 0.0001

Le résultat s'énonce de la manière suivante : à caractéristiques socio-démographiques (sexe et âge) identiques, à même milieu social et à niveaux d'entrée en 6ème comparables, les élèves en zone d'éducation prioritaire ont une probabilité plus élevée que les autres de passer en seconde générale : l'écart est de 6,6 points. L'effet marginal de la variable ZEP est somme toute relativement modeste.

II.5 Bilan d'étape

Le moment est venu de tirer enseignement des sections II.3 et II.4. Pour commencer, la table 9 donne les effets marginaux de la variable `zep` avec les différentes spécifications du modèle qui ont été utilisées. Pour la remplir, il suffit de passer la séquence des opérations écrite dans la section II.4, en redéfinissant à chaque fois la macro-variable `&listvar`. Par exemple, la première ligne de la table 9 donne la valeur de l'effet marginal de `zep` avec uniquement la variable *zep* dans le modèle. L'instruction correspondante définissant `&listvar` est :

```
%let listvar=zep;
```

Notons qu'on retrouve le résultat établi plus généralement page 44 : l'effet marginal de la variable `zep` est exactement égal à la différence (observée dans la table 1 page 55) de la part des élèves orientés en seconde générale entre ZEP ($x_1 = 1$) et hors ZEP ($x_1 = 0$). Les autres lignes de la table sont obtenues en ajoutant successivement les variables.

Table 9. Effet marginal de l'éducation prioritaire selon le modèle estimé

Modèle	Effet marginal	Écart-type
variable zep uniquement	−13,51***	1,34
zep + âge	−7,75***	1,26
zep + âge + sexe	−7,66***	1,25
zep + âge + sexe + csp	−1,12	1,12
zep + âge + sexe + csp + fran + math	6,62***	0,92
zep + âge + sexe + csp + fran + math + acad	5,04***	0,94

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

La table 9 illustre bien l'ambiguïté (et le terme est faible) de l'expression « toutes choses égales par ailleurs », qui est trop souvent prononcée mécaniquement lors des commentaires sur les résultats de l'estimation d'un modèle logit. Dans notre cas de figure, on pourrait très bien annoncer, selon le modèle retenu, « *toutes choses égales par ailleurs, étudier en ZEP a un impact négatif sur le passage en seconde générale* » (3ème ligne de la table), ou bien « *toutes choses égales par ailleurs, étudier en ZEP n'a pas d'impact sur le passage en seconde générale* » (4ème ligne), ou encore « *toutes choses égales par ailleurs, étudier en ZEP a un impact positif sur le passage en seconde générale* » (5ème ligne). Certes, il est logique de se fier à un modèle plus riche en variables. Mais si on n'avait pas disposé, dans notre source de données, d'information sur le niveau de l'élève en 6ème, on s'en serait probablement tenu à l'absence d'impact, « toutes choses égales par ailleurs ». Il est donc crucial de préciser ce que sont ces choses, et de rappeler que la conclusion pourrait changer si la source

de données contenait d'autres informations susceptibles d'influer sur l'orientation et introduisant un *effet de structure* supplémentaire.

Lorsqu'on examine la table, on est enclin à sélectionner deux moments : (1) celui où, en ajoutant la variable de milieu social, l'impact de l'éducation prioritaire devient non statistiquement significatif (même s'il reste négatif) ; (2) et celui où l'ajout des variables de niveau scolaire en 6ème fait changer le signe de l'impact. Cela ne permet pas d'affirmer que ces deux variables jouent les premiers rôles. On verra dans la section suivante quels outils mobiliser pour classer les variables selon leur ordre d'importance. Regardons tout de même de plus près leur impact.

On reprend les estimations en spécifiant un modèle qui fait dépendre l'orientation en seconde uniquement de la variable `zep` et de la variable de milieu social :

```
%let listvar=zep csp1 csp2 csp3 csp4 csp5 csp6;
```

On exécute ensuite la procédure `logistic`. Le paramètre de la variable `zep` vaut -0.216, qui se traduit par un effet marginal de $-4,38$ (avec un écart-type de $1,19$). L'introduction de la seule variable de milieu social fait donc passer l'écart, entre les élèves de ZEP et les autres, dans le taux de passage en seconde générale, de $-13,5\%$ à $-4,4\%$. Autrement dit, la différence de milieu social entre les élèves de ZEP et les autres explique à elle seule plus des deux-tiers (67,4% pour être précis [41]) de l'écart « brut » constaté de 13,5%. Le milieu social joue bien un rôle important.

Cette manière de présenter les résultats est analogue à ce que produit la *décomposition d'Oaxaca-Binder*, selon l'appellation consacrée. Cette méthode de décomposition a été présentée à l'origine par les deux auteurs dans deux publications différentes [42], appliquée à la discrimination salariale, entre hommes et femmes notamment. Le principe est de décomposer l'écart salarial entre les hommes et les femmes en une partie expliquée par les caractéristiques observées introduites dans l'analyse, et une partie résiduelle, c'est-à-dire restant inexpliquée. De la même manière, en revenant à notre sujet, 71,2% de l'écart dans l'orientation en seconde est expliqué par la catégorie sociale des parents, et les 28,8% restants expliqués par d'autres facteurs.

Second exercice, on fait dépendre l'orientation uniquement du niveau de l'élève en 6ème (en plus, bien entendu, de la ZEP) :

```
%let listvar=zep fran math;
```

Le paramètre de la variable `zep` est positif : $0,205$ (écart-type de $0,068$). Converti en effet marginal, il vaut $3,13$ points (écart-type de $1,02$). On peut spécifier un modèle encore plus parcimonieux, en définissant une variable de niveau qui cumule le français et les mathématiques (`niveau=fran+math`) :

41. $(13,5 - 4,4)/13,5 = 0,674$

42. A.S. Blinder (1973), « Wage discrimination : reduced form and structural estimates », *Journal of Human Resources*, 8 (4) ; R.Oaxaca (1973), « Male-female wage differentials in urban labor markets », *International Economic Review*, 14 (3).

```
%let listvar=zep niveau;
```

Les résultats des estimations donnent $0,208$ comme valeur du paramètre de la variable `zep`, ce qui correspond à un effet marginal de $3,18$ points, valeurs très proches des précédentes.

Ainsi, à même niveau initial en 6ème, les élèves de ZEP ont en moyenne une probabilité de passer en seconde générale supérieure à celle des autres élèves. L'écart moyen est de $3,2$ points.

Ce résultat ne provient pas d'une lecture directe des informations collectées par le panel qui a suivi une cohorte d'enfants entrés en 6ème en 1995. Il est issu d'une modélisation, c'est-à-dire de la spécification d'un modèle – très simple puisqu'impliquant seulement les deux variables `zep` et `niveau` – dont les paramètres ont été estimés sur les données du panel 1995.

En fait, on peut faire une lecture plus directe des informations, sans passer par un modèle. On procède comme suit. On découpe notre population d'élèves en groupes de niveau, selon leurs résultats aux tests effectués en 6ème mesurés par la variable `niveau`. On a retenu ici 20 groupes. Le premier (*resp.* dernier) groupe rassemble les 5% d'élèves ayant eu les résultats les plus faibles (*resp.* les meilleurs). Dans chaque groupe, on calcule la proportion des élèves de ZEP qui sont passés en seconde générale et celle des élèves hors ZEP qui ont été orientés en seconde générale. On compare ensuite les deux proportions dans chacun des 20 groupes.

Concrètement, pour répartir la population en 20 groupes de taille équivalente, on écrit :

```
proc rank data=tab groups=20 out=tabg;
  var niveau;
  ranks pniveau;
run;
```

La table en sortie de la procédure (option `out=`), nommée ici `tabg`, est la copie conforme de la table `tab` augmentée d'une variable, nommé `pniveau`, qui identifie chacun des 20 groupes et qui prend les valeurs 0 à 19 (et non 1 à 20). On calcule ensuite, groupe par groupe, les proportions d'élèves passés en seconde générale en distinguant les élèves de ZEP et les autres :

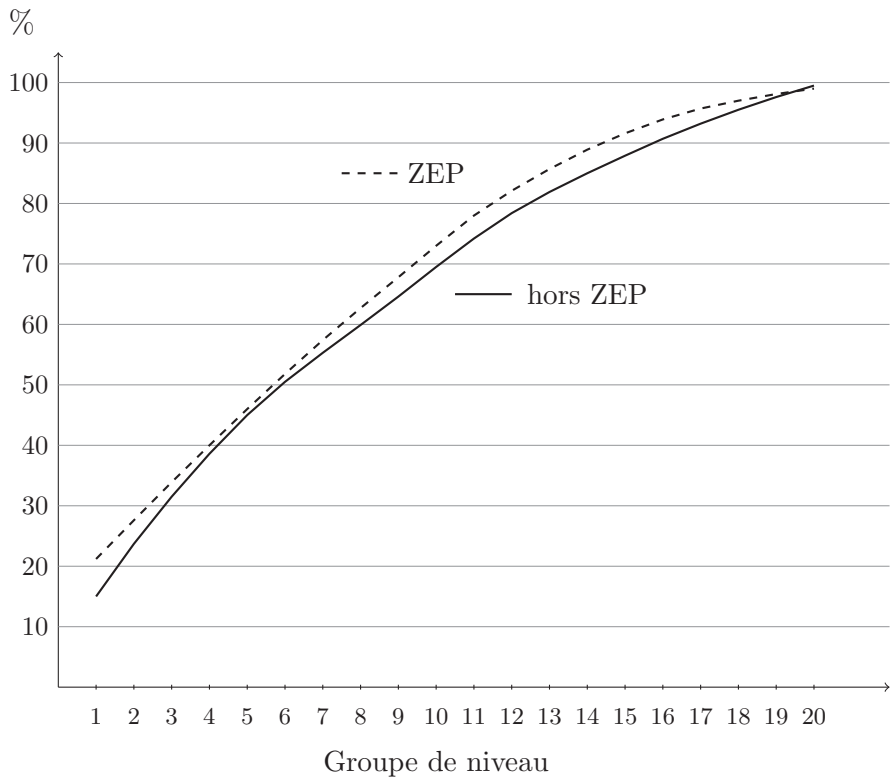
```
proc summary data=tabg nway;
  class pniveau zep;
  var secondeg;
  output out=tabs mean=prop_seconde;
run;
```

On récupère les statistiques souhaitées dans la table nommée ici `tabs`. La variable `prop_seconde` donne la proportion de passage en seconde générale pour chaque groupe (variable `pniveau`), selon l'appartenance ou non en ZEP (variable `zep`).

On représente ces proportions dans un plan avec, en abscisse, le groupe de niveau et, en ordonnée, les proportions de passage en seconde générale. On trace deux courbes, l'une reliant les proportions des élèves de ZEP et l'autre reliant celles des autres élèves. La courbe des élèves de ZEP étant un peu chahutée à cause des effectifs relativement faibles des groupes, on préfère représenter des courbes lissées [43]. Elles offrent un plus grand confort de lecture sans trahir les résultats.

La courbe représentant les taux de passage, en seconde générale, des élèves de ZEP est globalement au-dessus de celle des élèves hors ZEP (figure 1). L'écart entre les deux courbes varie autour de 3%, avec un maximum pour le premier groupe (6, 2 points), puis pour le groupe « médian » (4, 5 points), et un minimum pour les groupes les plus élevés. Le constat est donc tout à fait cohérent avec le résultat du modèle simple à deux variables, qui donne un écart moyen de 3, 2 points.

Figure 1. Taux de passage en seconde générale des élèves de ZEP et des élèves hors ZEP, selon leur groupe de niveau en 6ème



À ce stade, on peut se demander si la présentation des résultats de l'étude sur le rôle de l'éducation prioritaire dans l'orientation post-troisième ne devrait pas reposer essentiellement sur la figure 1. Celle-ci, en effet, ne s'appuie sur aucun modèle. Même

43. C'est la procédure *loess* de SAS qui a été utilisée ici. Des outils de lissage sont disponibles sous Excel.

si les données ont été un peu « travaillées » (découpage de la population des élèves en 20 groupes, lissage des courbes), il s'agit avant tout de la représentation graphique de statistiques descriptives, représentation qui, davantage qu'un modèle et ses résultats, est susceptible de « marquer les esprits ». De plus, un modèle logit plus complet (sections II.3.d ou II.3.e) apporte une valeur ajoutée toute relative : il estime à 5 ou 6 points (voir section II.4) la différence moyenne dans les taux de passage au lieu des 3 points de la figure 1, ce qui ne bouleverse pas la donne.

Cela étant, la modélisation a été et reste utile. D'abord, il n'était pas *a priori* évident que l'essentiel des résultats pouvait se résumer à la figure 1. Ceci est apparu à l'issue de la démarche de modélisation. Ensuite, les résultats du modèle dans sa plus simple expression (avec seulement les variables `zep` et `niveau`) permettent d'une part de chiffrer précisément l'écart moyen des deux courbes (3, 2 points), d'autre part de pouvoir affirmer que cet écart est statistiquement significatif (au seuil de 1%). Autrement dit, la différence est réelle et ne repose pas sur les aléas de l'échantillon.

Un dernier mot. À dire vrai, et en première lecture, on peut estimer que les deux résultats qu'on vient d'établir sont incohérents. D'un côté, la dimension sociale explique 70% de l'écart dans l'orientation post-troisième entre les élèves de ZEP et les autres (voir *supra*, page 76). Elle capterait donc la plus grande partie des *effets de composition* des ZEP sur l'orientation post-troisième. De l'autre côté, prendre en compte uniquement le niveau à l'entrée en sixième rend positif l'impact de l'éducation prioritaire sur la probabilité de passer en seconde générale. La différence de structure des populations en ZEP et hors ZEP serait d'abord une différence de niveau initial.

En réalité, il n'y a pas de contradiction. Niveau initial de l'élève et catégorie sociale de ses parents sont évidemement liés. Ainsi, lorsque le modèle ne retient comme variable de contrôle que le milieu social pour conclure que ce dernier explique 70% de l'écart d'orientation entre élèves de ZEP et hors ZEP, cette variable « embarque » aussi avec elle la différence de niveau scolaire des enfants appartenant à des milieux différents. Et on ne saurait dire laquelle des deux variables – niveau scolaire en 6ème et milieu social – a la prééminence sur l'autre pour capter les effets de structure.

II.6 Changement de perspective (I) – Qu'est-ce qui distingue les élèves s'orientant en seconde générale ?

Tout en restant sur les mêmes données et les mêmes variables[44], on change de perspective pour montrer l'autre aspect du modèle *logit* : l'outil d'analyse discriminante. On ne centre plus l'analyse sur le rôle spécifique de l'éducation prioritaire dans l'orientation post-troisième. On souhaite maintenant aborder la problématique suivante.

De manière très (trop ?) schématique, deux opinions s'opposent sur les déterminismes à l'œuvre dans les destins scolaires des collégiens. La première consiste à dire que les choses se jouent en grande partie au cours du primaire, que la suite de la scolarité est largement déterminée par le niveau atteint en fin de CM2. Le second discours insiste lui sur le rôle déterminant de la famille lors des études secondaires. Les inégalités d'orientation sont le reflet des inégalités sociales. Un parent de milieu favorisé a davantage de ressources – financières, intellectuelles, ... – pour accompagner ses enfants sur le chemin de la réussite. Bien entendu, et on le dira ultérieurement, la situation est plus complexe que cela. Partons néanmoins de ces deux positions tranchées.

La question est : laquelle des deux dimensions – milieu social et niveau de l'élève en 6ème – joue le premier rôle dans l'orientation en fin de collège ? Dans cette perspective, il n'y a plus de variable principale. Toutes les variables sont mises sur le même plan, même si on en privilégie *a priori* deux pour les besoins de l'analyse.

Pour répondre à la question, il faut « décorréler » les variables. En effet, quand on compare les taux de passage en seconde générale des enfants de cadres (91,3%) et des enfants d'ouvriers (52,4%), on pense tenir là un facteur de distinction de première importance puisque quelque 40 points (38,9 pour être précis) les séparent. Or, on constate que, d'une manière générale, les élèves les mieux notés en 6ème poursuivent plus fréquemment que les autres leurs études dans la voie générale ou technologique : ceux passés en seconde générale avaient obtenu en moyenne 57,3 points aux épreuves de mathématiques de 6ème, contre 44,9 points pour les autres. Il se trouve que les enfants de cadres ont eu une meilleure moyenne (59,4 points *vs* 44,3 points pour les enfants d'ouvriers) à ces épreuves. Par conséquent, le fait que les enfants de cadres et d'ouvriers ne connaissent pas la même orientation à la fin du collège s'explique au moins en partie par leur niveau à l'entrée en 6ème. Reste à savoir si cette part expliquée est faible ou importante.

Décorréler les variables, c'est adopter la démarche analytique au cœur du modèle *logit*, qui, en s'appuyant sur l'hypothèse d'additivité, permet d'estimer le rôle propre joué par chaque variable (page 10) dans l'orientation post-troisième. On ne se limitera pas à introduire seulement les variables de milieu social et de niveau de l'élève, pour deux raisons. La première est qu'on a privilégié ces deux dimensions en sup-

44. Le parti de travailler sur les mêmes variables a été pris par souci de continuité dans l'exposé. Il se révèlera critiquable (voir fin de la section II.7.a).

posant *a priori* qu'elles étaient les plus importantes dans le processus d'orientation. C'est une hypothèse qui doit être vérifiée. Pour ce faire, il faut introduire d'autres variables. Deuxième raison, on a vu (section II.3.c) que la proportion de filles n'était pas exactement la même d'une catégorie sociale à une autre. Par ailleurs, les filles ont eu en moyenne, par rapport aux garçons, de meilleurs résultats aux tests de français (mais pas aux tests de mathématiques). Le facteur *Sexe de l'élève* étant lié à la fois au milieu social (même s'il l'est faiblement) et au niveau en 6ème (même s'il l'est de manière complexe), on a intérêt à l'isoler pour mieux mettre en balance les deux dimensions qui nous intéressent de prime abord.

Tout compte fait, on reproduit les estimations du modèle de la section II.3.d. Notons au passage qu'il est préférable de ne pas introduire la dimension académique comme dans la section II.3.e, car les indicatrices qui la représentent ne nous disent pas précisément ce qui est mesuré.

Cela étant, le problème avec la table 8 est que les valeurs des paramètres sont peu parlantes. Pour une meilleure lisibilité, on a intérêt à les transformer en points de pourcentage, en calculant les effets marginaux des variables auxquelles les paramètres sont associés.

On va donc étendre aux autres variables le calcul de l'effet marginal effectué section II.4 pour la variable `zep`, qui est une variable binaire. L'extension à une variable polytomique, comme la variable d'âge à l'entrée en sixième ou la variable de milieu social, se fait sans problème, comme on le verra ultérieurement. La difficulté provient des variables de scores aux épreuves de français et de mathématiques, qui sont continues et en se comptent pas en unité de mesure (voir page 45). Pour calculer des effets marginaux au même titre que les autres variables, il faut d'abord les transformer en variables polytomiques. Pour ce faire, et pour chacune des variables `fran` et `math`, on a choisi de distinguer quatre groupes d'élèves selon leur position par rapport aux quatre quartiles de la distribution du score. Pour le score en français par exemple, on aura donc un premier groupe d'élèves rassemblant les 25% ayant eu les moins bons résultats aux tests, un deuxième groupe comprenant les 25% suivants dans l'ordre croissant des résultats, un troisième constitué des 25% suivants, les 25% ayant eu les meilleurs résultats faisant partie du dernier groupe. On transforme la variable continue `fran` en une variable polytomique ordonnée à quatre modalités. On a intérêt à prendre comme référence la modalité correspondant au premier quartile (voir page 23). Notons que le choix de 4 groupes est arbitraire, on pourrait en définir 5 en répartissant les élèves selon les quintiles de la distribution, ou bien 10 en retenant les déciles.

Pour définir les quatre groupes, on utilise la procédure `rank` de SAS. Pour le test en français (variable `fran`), la syntaxe en est la suivante :

```
proc rank data=tab groups=4 out=tab;
  var fran;
  ranks qfran;
run;
```

La procédure crée quatre groupes[45] (option `groups=4`) numérotés de 0 à 3 (et non de 1 à 4) par la variable `qfran`. La table en sortie, qu'on a choisie identique à la table en entrée, est enrichie de la variable `qfran`. On réalise le même exercice avec la variable `math`. On crée ensuite les quatre variables binaires représentant les quatre modalités de `qfran` et `qmath` :

```
data tab;
  set tab;
  array qfr(i) qfr1-qfr4;
  array qma(i) qma1-qma4;
  do i=1 to 4;
    qfr=(i=qfran+1);
    qma=(i=qmath+1);
  end;
run;
```

Puis on enchaîne les instructions suivantes. On définit d'abord, par une macro-variable, la liste des variables du modèle :

```
%let listvar1=zep retard avance fille csp1 csp2 csp3 csp4 csp5 csp6
              qfr2 qfr3 qfr4 qma2 qma3 qma4;
```

avant d'exécuter la procédure `logistic` :

```
proc logistic data=tab descending covout outest=param1;
  model secondeg = &listvar1;
run;
```

puis la macro `marginal` :

```
%marginal(tab_ent=tab,x=&listvar1,
          param_ent=param1,var_qual=qfr2 qfr3 qfr4);
```

et :

```
%marginal(tab_ent=tab,x=&listvar1,
          param_ent=param1,var_qual=qma2 qma3 qma4);
```

Attention de lister les 3 modalités de chaque variable (rappel : la modalité de référence est exclue) dans l'ordre où elles apparaissent lors de la définition de la macro-variable `listvar1`. Tout ceci donne en sortie :

Effet marginal de 'qfr2 qfr3 qfr4'

	effet marginal	écart_type	significativité
QFR2	10.8619	1.0556	< 0.0001
QFR3	16.9216	1.2141	< 0.0001
QFR4	27.4096	1.3862	< 0.0001

45. Tous les élèves notés identiquement étant affectés au même groupe, les quatre groupes n'ont pas exactement le même effectif.

et :

Effet marginal de 'qma2 qma3 qma4'

	effet marginal	écart_type	significativité
QMA2	11.9175	1.0984	< 0.0001
QMA3	19.9986	1.2125	< 0.0001
QMA4	32.9009	1.3335	< 0.0001

Dernier exemple, la variable d'âge à l'entrée en sixième :

```
%marginal(tab_ent=tab,x=&listvar,
          param_ent=param,var_qual=retard avance);
```

On obtient en sortie :

Effet marginal de 'retard avance'

	effet marginal	écart_type	significativité
RETARD	-21.6386	1.1573	< 0.0001
AVANCE	15.8242	2.1166	< 0.0001

La table 10 rassemble tous les éléments. La première colonne reprend les valeurs estimées des paramètres, la deuxième les traduit en points de pourcentage (effets marginaux) et la troisième donne les écarts bruts. Les écarts bruts mesurent simplement les différences constatées des taux de passage entre chaque modalité d'une variable et sa modalité de référence. Ainsi, on retrouve pour la variable ZEP l'écart de 13,5 points reporté dans la table 1 (page 55). Autre exemple, la différence de taux de passage entre les élèves en retard et ceux à l'heure en 6ème s'établit à 43,1 points, au détriment des premiers. Les effets marginaux (deuxième colonne) pourraient aussi être nommés écarts résiduels. L'écart résiduel mesure le rôle joué en propre par chaque variable (le rôle restant à chaque variable, si on préfère) lorsque les autres variables sont maintenues constantes.

La comparaison des colonnes 2 et 3 permet d'apprécier le changement induit par la « décorrélation » des variables ou dimensions. Concernant notamment les variables de milieu social et de niveau de l'élève en 6ème, on obtient les résultats suivants. Pour les enfants de cadres, l'écart brut, c'est-à-dire la différence constatée entre le taux de passage en seconde générale de ces enfants et le taux des enfants des familles dont le chef a déclaré être sans activité professionnelle ou n'a pas répondu à la question (population de référence), s'établit à 51,8 points. Pour les enfants d'ouvriers, il vaut 12,9 points. Par conséquent, l'écart brut entre enfants de cadres et enfants d'ouvriers est de 38,9 points ($51,8 - 12,9$). Quand on passe aux écarts résiduels (effets marginaux), ils valent respectivement pour les enfants de cadres et pour les enfants d'ouvriers 26,5 points et 4,7 points (toujours par rapport à la

Table 10. Les résultats du modèle

	Paramètre	Effet marginal (%)	Écart brut (%)
Age à l'entrée en sixième (*ref=à l'heure*)			
en retard	−1,321***	−21,6	−43,1
en avance	1,229***	15,8	22,0
Sexe de l'élève (*ref=garçon*)			
fille	0,532***	7,7	7,8
Milieu social de l'élève (*ref=sans act., nr*)			
agriculteurs exploitants	0,431**	7,1	27,0
artis., commerç., chefs d'entrep.	0,685***	11,1	29,3
cadres, prof. intell. sup.	1,842***	26,5	51,8
prof. intermédiaires	1,053***	16,6	39,3
employés	0,528***	8,7	22,9
ouvriers	0,281**	4,7	12,9
Niveau en français (*ref=groupe 1*)			
groupe 2	0,610***	10,9	27,8
groupe 3	0,984***	16,9	45,1
groupe 4	1,759***	27,4	59,9
Niveau en maths (*ref=groupe 1*)			
groupe 2	0,645***	11,9	25,7
groupe 3	1,133***	20,0	43,7
groupe 4	2,146***	32,9	58,8
Appartenance à une ZEP (*ref=non*)			
oui	0,389***	5,4	−13,5

Seuils de significativité : *** = 1% ; ** = 5% ; * = 10%.
Source : DEPP – Panel 1995.

référence). En conséquence, l'écart résiduel entre enfants de cadres et d'ouvriers s'établit à 21,8 points $(26,5 - 4,7)$, soit 17 points de moins que l'écart brut.

On remarquera que la valeur de l'effet marginal de `zep` n'est pas la même que celle calculée avec le modèle spécifié pourtant avec les mêmes variables (page 73). À ceci près que ce n'est pas exactement les mêmes variables. Certes, on retrouve dans l'un et l'autre cas l'âge à l'entrée en sixième et le sexe de l'élève, son milieu social, son niveau en français et en mathématiques, l'appartenance de son établissement à une ZEP. Mais dans le premier cas ce sont les variables continues mesurant les niveaux en français et en mathématiques qui ont été introduites, alors que dans cette section ce sont des groupes de niveau qui ont été retenus. Cela étant, les deux valeurs de l'effet marginal ne sont pas significativement différentes compte tenu de leurs écarts-types respectifs.

Une dernière remarque. On aurait pu découper les variables `fran` et `math` en 20 groupes au lieu de 4, c'est-à-dire classer les élèves en groupes de niveau selon les vingtiles des distributions des scores en français et en mathématiques au lieu

des quartiles, de manière à capter plus finement l'impact du niveau des élèves, dans l'hypothèse où l'impact varierait en fonction du niveau même de l'élève. De fait, lorsqu'on réestime le modèle avec les 20 groupes en français et en mathématiques, l'indicateur d'Akaïke (voir section I.4) vaut 11 842 au lieu de 11 974 pour le modèle avec 4 groupes. Cette valeur plus faible est le signe d'un modèle de meilleur qualité. Mais si on examine le critère de Schwartz, il passe de 12 101 pour le modèle à 4 groupes de niveau à 12 211 pour celui à 20 groupes, signe cette fois-ci d'une dégradation de la qualité du modèle. Il faut se rappeler que le critère de Schwartz pénalise davantage que le critère d'Akaïke les modèles peu parcimonieux. Il est donc plus sensible à cette « inflation » de variables créées par les 20 groupes de niveau. Il est donc important de limiter le nombre de variables, en tout cas de ne pas introduire de variables polytomiques avec un nombre démesuré de modalités, qu'il faut donc préalablement regrouper.

II.7 Changement de perspective (II) – Quelle hiérarchie des variables ?

La table 10 confirme que toutes les variables retenues sont discriminantes. L'étape suivante est de déterminer celles qui jouent les premiers rôles dans l'orientation post-troisième. On cherche ainsi à établir une hiérarchie des variables par ordre d'importance.

Les résultats des estimations du modèle figurant dans la table 10 restent insuffisants pour réaliser l'exercice. Certes, on peut classer deux variables binaires par ordre d'importance en comparant les valeurs (absolues) de leurs paramètres ou de leurs effets marginaux. On conclura ainsi que le sexe de l'élève joue un rôle plus important (il est plus discriminant) que l'appartenance à une zone d'éducation prioritaire. Pour les variables polytomiques, on peut toujours comparer les amplitudes des paramètres ou des effets marginaux. Par exemple, pour la variable de milieu social, les effets marginaux vont de 0 (pour la modalité de référence, par définition) à 26,5 pour la modalité *cadres et professions intellectuelles supérieures*. L'amplitude des effets marginaux est donc de 26,5. Pour la variable d'âge à l'entrée en sixième, l'effet marginal le plus faible est celui de la modalité *en retard* (-21,6) et le plus élevé celui de la modalité *en avance* (15,8). L'amplitude est de 37,4. L'âge serait ainsi plus discriminant que le milieu social. Mais la conclusion reste incertaine. Surtout, cette manière de faire ne permet pas de régler le cas des variables continues qui ont été transformées en variables polytomiques ordonnées, comme ce qui a été fait avec nos deux variables de niveau en français et en mathématiques, découpées en 4 groupes. L'amplitude des effets marginaux est de 32,9 pour les mathématiques et de 27,4 pour le français (table 10). L'âge d'entrée en sixième – amplitude de 37,4 – jouerait donc un rôle plus important que le niveau en mathématiques ou en français de l'élève en 6ème. Mais si on transformait nos deux variables continues en variables polytomiques à 10 modalités (selon les déciles des distributions des scores), alors l'amplitude des effets marginaux serait de 44,0 pour la variable de niveau en mathématiques et de 32,5 pour celle en français. Dans ce cas, l'âge à l'entrée au collège passerait derrière le niveau en mathématiques à l'entrée en 6ème.

Il faut donc se tourner vers une autre méthode. Mais, à notre connaissance, il n'en existe pas qui soit théoriquement éprouvée. Celle proposée ici est de nature heuristique. Elle s'appuie sur des indicateurs de qualité du modèle (voir section I.4).

II.7.a Utilisation d'un critère de prédiction

Pour classer des variables selon leur importance, un premier moyen est d'utiliser le pseudo-R^2 proposé par Wooldridge (voir section I.4.b, page 32), qui mesure le pouvoir prédictif du modèle, c'est-à-dire sa capacité à prédire l'appartenance à l'une ou l'autre des catégories C_0 ou C_1, compte tenu des variables x_k. *A priori*, plus on dispose d'information sur l'individu, c'est-à-dire plus le nombre de variables est élevé, mieux on saura prédire son appartenance à C_0 ou C_1. *A priori* donc, le pseudo-R^2 augmente avec le nombre de variables x_k. *A contrario*, si on supprime des variables

du modèle, il perd en capacité prédictive et le pseudo-R^2 diminue.

La démarche est alors la suivante. On part du modèle de la section II.3.d, considéré comme complet. Formellement, les variables du modèle sont au nombre de 12, si on comptabilise toutes les indicatrices associées aux modalités des variables polytomiques. Elles peuvent être regroupées en 5 dimensions : l'âge de l'élève à son entrée en 6ème (représenté par les variables binaires `retard` et `avance`), son sexe, son niveau au début du collège (mesuré par les deux variables continues `fran` et `math`), sa scolarisation ou non dans une ZEP, son milieu social (variables `csp1` à `csp6`).

On calcule le pseudo-R^2 du modèle complet. Puis on supprime une des dimensions (la dimension ZEP par exemple). On estime le modèle ainsi réduit et on en déduit le pseudo-R^2. On repart du modèle complet, dont on enlève une des 3 autres dimensions (l'âge par exemple). On estime le modèle obtenu et on note son pseudo-R^2. Et ainsi de suite. La dimension la plus influente est celle qui, lorsqu'on la retire du modèle, dégrade le plus la qualité prédictive du modèle, c'est-à-dire provoque la plus forte baisse du pseudo-R^2. Les dimensions sont ainsi classées selon l'écart entre le pseudo-R^2 du modèle complet et celui calculé avec le modèle sans la dimension considérée.

La procédure `logistic` ne produit pas automatiquement la valeur du pseudo-R^2. Il faut écrire des instructions spécifiques. On commence par calculer le pseudo-R^2 du modèle complet. Pour ce faire, on part de la liste des variables du modèle, nommée `listvar` et définie par :

```
%let listvar=zep retard avance fille csp1 csp2 csp3 csp4 csp5 csp6
             fran math ;
```

Puis on exécute la procédure `logistic` avec l'instruction `output` :

```
proc logistic data=tab descending noprint;
  model secondeg=&listvar;
  output out=p0 p=pred0;
run;
```

L'instruction `output` crée une table, appelée `p0`, qui est l'image de la table en entrée, `tab`, augmentée de la variable nommée ici `pred0`. Cette variable, créée par option `p=` de l'instruction `output`, est la probabilité prédite par le modèle – et notée $\hat{P}(y_i = 1|x_i)$ – que l'individu i passe en seconde générale. On est alors en mesure de calculer la corrélation des y_i et des $\hat{P}(y_i = 1|x_i)$, qui est précisément le pseudo-R^2 recherché (voir section I.4.b). Pour ce faire, on utilise la procédure `corr` :

```
proc corr data=p0;
  var secondeg pred0;
run;
```

ce qui donne :

```
Pearson Correlation Coefficients, N = 13499
          Prob > |r| under H0: Rho=0

                                secondeg         pred0

secondeg                         1.00000       0.58891
                                               <.0001

pred0                            0.58891       1.00000
Estimated Probability             <.0001
```

La corrélation s'établit à 0,589. Le pseudo-R^2 est égal à son carré, soit 0,347. On calcule ensuite le pseudo-R^2 du modèle réduit qui est obtenu en supprimant la variable ZEP du modèle complet. Comme précédemment, on exécute successivement les deux procédures `logistic` et `corr` mais en remplaçant la liste des variables `&listvar` par la liste `&listvar1` définie par :

```
%let listvar1=retard avance fille csp1 csp2 csp3 csp4 csp5 csp6
              fran math ;
```

On fait de même avec le modèle issu du modèle complet mais sans la dimension d'âge à l'entrée en 6ème. On en chaîne les procédures `logistic` et `corr` en utilisant la liste des variables `listvar2` :

```
%let listvar2=zep fille csp1 csp2 csp3 csp4 csp5 csp6
              fran math ;
```

Et ainsi de suite.

La table 11 présente les résultats de l'exercice. La première ligne donne la valeur du pseudo-R^2 pour le modèle complet. Chaque ligne suivante donne la valeur obtenue lorsqu'on enlève alternativement une seule des 5 dimensions.

C'est quand on exclut les variables mesurant le niveau de l'élève qu'on perd le plus d'information. De ce point de vue, le niveau de l'élève est donc la plus importante des 5 dimensions. Viennent ensuite l'âge de l'élève à l'entrée en sixième, son milieu social, son sexe et la scolarisation en ZEP.

À première vue, et contrairement à ce qu'on avait présupposé, le milieu social ne ferait pas partie des deux dimensions les plus importantes dans le processus d'orientation en fin de troisième. Il viendrait après le niveau en 6ème et après l'âge auquel l'enfant est entré au collége. Mais, à l'analyse, ce constat est très fragile.

D'abord, la variable d'âge mesure, du moins partiellement, le niveau atteint par l'élève à la fin du primaire. L'entrée retardée au collège est, en effet, la conséquence d'un redoublement au cours des années précédentes et donc le signe de difficultés scolaires rencontrées par l'enfant. Dans ces conditions, l'âge apparaît comme une variable redondante, le niveau de l'élève étant mieux mesuré par les tests de français et de mathématiques. En d'autres termes, le modèle se révèle être mal spécifié, mal

Table 11. Qualité du modèle selon les dimensions exclues (selon le pseudo-R^2)

	pR^2
Modèle complet	0,347
ZEP exclue	0,345
âge en sixième exclu	0,319
sexe exclu	0,340
milieu social exclu	0,321
niveau en sixième exclu	0,194

Lecture : lorsqu'on exclut du modèle les variables caractérisant le milieu social de l'élève, le pseudo-R^2 vaut 0,321.
Source : DEPP – Panel 1995.

adapté finalement au débat engagé au début de la section II.6 sur la prééminence du milieu social ou du niveau de l'élève. Par souci de clarté, il aurait peut-être été préférable de ne pas retenir l'âge [46].

Ensuite, même en supposant qu'on ait su classer sans ambiguïté le niveau à l'entrée en 6ème devant le milieu social, cela ne permet pas de conclure que le premier prime sur le second. Car le niveau en 6ème est lui-même le résultat d'apprentissages antérieurs qui sont socialement marqués, si bien que cette variable de niveau contient en elle-même, si on peut dire, une part de « social » due au rôle joué par le contexte familial *avant* l'entrée en 6ème. Ceci interdit donc de faire clairement la part des choses. En d'autres termes, avec les données à notre disposition, nous ne pouvons pas répondre à la question sur la prééminence, dans l'absolu, des rôles respectifs du niveau de l'élève à son entrée au collège et de son milieu social [47]. Le rôle du milieu social mis en évidence par le modèle est à tout le moins celui limité aux années collège.

II.7.b Utilisation d'un critère d'information

Une deuxième possibilité pour hiérarchiser les variables est d'utiliser un critère d'information (voir section I.4.a, page 30), en utilisant une démarche analogue à la précédente.

On part du modèle complet, celui contenant toutes les variables. Lorsqu'on supprime une ou plusieurs variables x_k, le modèle ainsi réduit s'éloigne encore davantage de la réalité que le modèle complet. On perd donc de l'information, et les critères d'information d'Akaïke ou de Schwartz augmentent (voir section I.4.a). L'augmentation est d'autant plus forte que la variable (ou les variables) supprimée(s) joue(nt)

46. Ceci illustre la nécessité de bien réfléchir aux variables à introduire dans un modèle pour répondre à une question donnée.
47. ... pour autant que cette question ait un sens !

un rôle important dans l'adéquation du modèle à la réalité.

Appliquons ce principe au modèle de la section II.3.d, où les variables ont été regroupées en 5 dimensions comme dans la section précédente. Le modèle complet fournit un certain niveau d'information sur le processus de passage en seconde générale. Ce niveau est mesuré par le critère d'Akaïke AIC ou le critère de Schwartz SC, dont les valeurs sont fournies automatiquement par la procédure `logistic` (voir la colonne `Intercept and Covariates` du bloc intitulé `Model Fit Statistics`, page 60) et que l'on note. Puis on supprime une des 5 dimensions du modèle, on estime le modèle réduit ainsi obtenu, et on note les nouvelles valeurs des critères. On repart ensuite du modèle complet, on supprime une des autres dimensions, on estime le modèle qui s'en déduit, et on récupère les valeurs de AIC ou SC. Et ainsi de suite jusqu'à avoir retiré alternativement toutes les dimensions. Celle dont la suppression provoque la plus forte perte d'information, c'est-à-dire la plus forte hausse des critères AIC ou SC, sera alors considérée comme la plus importante.

La table 12 rassemble les résultats de l'exercice. La première ligne donne les valeurs des critères SC et AIC pour le modèle complet. Chaque ligne suivante donne les deux valeurs obtenues lorsqu'on enlève une seule des 5 dimensions.

Table 12. Qualité du modèle selon les dimensions exclues (critères d'information de Schwartz et d'Akaïke)

	SC	*AIC*
Modèle complet	12 023,3	11 925,6
ZEP exclue	12 060,9	11 970,8
âge en sixième exclu	12 437,7	12 355,1
sexe exclu	12 133,1	12 043,0
milieu social exclu	12 396,4	12 343,8
niveau en sixième exclu	14 402,8	14 320,2

Lecture : lorsqu'on exclut du modèle les variables caractérisant le milieu social de l'élève, les critères de Schwartz et d'Akaïke valent respectivement 12 396,4 et 12 343,8.

Source : DEPP – Panel 1995.

On retrouve la même hiérarchie des dimensions que précédemment avec un critère de prédiction.

II.8 La question des pondérations

L'échantillon d'étude peut ne pas être représentatif de la population générale pour deux raisons :

- parce que – cause en amont de l'enquête – lors du tirage de l'échantillon, certaines catégories ont été surreprésentées (par exemple, les élèves d'établissement en éducation prioritaire) : l'échantillon a été constitué avec un tirage à probabilités inégales ;
- parce que – cause en aval de l'enquête – tous les enquêtés n'ont pas répondu, et ceux qui ont échappé à l'enquête sont particuliers si bien que l'échantillon des répondants n'est pas représentatif de l'ensemble de la population.

Lorsque l'échantillon n'est pas représentatif pour l'une ou l'autre raison, alors il faut pondérer les observations individuelles de manière à reconstituer un échantillon à l'image de la population générale. Obtenir le bon jeu de pondérations est plus ou moins simple, selon la cause de non représentativité.

Si elle se situe exclusivement en amont, c'est-à-dire si elle est entièrement imputable au plan de sondage, le redressement est aisé à faire. Les poids sont calculés avec l'inverse de la probabilité de tirage. Par exemple, si les élèves en éducation prioritaire ont été tirés avec une probabilité double de celle des autres élèves, ils seront proportionnellement deux fois plus nombreux dans l'échantillon que dans la population générale. Ils devront alors peser deux fois moins dans l'échantillon pour que celui-ci retrouve sa représentativité.

Si la cause se situe en aval de l'enquête, si elle tient à la spécificité des répondants, alors le redressement peut être très délicat à réaliser, surtout si on suspecte que les répondants se sont autosélectionnés sur des caractéristiques inobservées dans l'enquête.

Supposons qu'on dispose d'un jeu de pondérations affectées aux individus de l'échantillon. Faut-il les utiliser pour estimer correctement les paramètres du modèle ? Cette question est moins simple qu'on ne le pense *a priori*[48]. Insistons d'abord sur un point pratique : il faut vérifier que la somme des poids utilisés soit égale à la taille de l'échantillon (on dit alors que les poids sont normalisés). Sinon, les écarts-types des différents paramètres obtenus en pondérant les observations seront biaisés, avec le risque de conduire à des conclusions fortement erronées sur leur significativité statistique.

Notons $\tilde{X}$ l'ensemble des variables qui ont été éventuellement utilisées, d'une part pour stratifier l'échantillon et faire un tirage à probabilités inégales, d'autre part pour traiter la non-réponse. Supposons dans un premier temps que la non-réponse ait été correctement corrigée, c'est-à-dire qu'on n'ait pas oublié dans la liste $\tilde{X}$ de variables distinguant les répondants des non répondants. Si toutes les variables $\tilde{X}$ sont introduites dans le modèle *logit*, si elles se retrouvent toutes dans

48. Pour son traitement complet, voir L. Davezies et X. D'Haultfœuille (2009), « Faut-il pondérer ? . . . Ou l'éternelle question de l'économètre confronté à des données d'enquête », *Document de travail de la Direction des Études et Synthèses Économiques*, Insee, n° 2009/06.

la liste de variables x du modèle, alors la question de pondérer ou non n'a pas d'importance : on obtient dans les deux cas des estimations sans biais. S'il fallait choisir, on opterait plutôt pour ne pas pondérer, car dans ce cas les estimations obtenues sont plus précises. En revanche, si x ne contient pas toutes les variables corrigeant la sélection, alors il faut pondérer sinon les estimations des paramètres sont, en règle très générale, biaisées.

Supposons maintenant que la liste $\tilde{X}$ ne soit pas complète, que, par exemple, le concepteur d'enquête ait redressé la non-réponse sur un nombre insuffisant de variables. Si on pense que la liste x est, elle, complète, c'est-à-dire que le redressement aurait été correct en l'utilisant, alors il n'est pas important de pondérer. Toutefois, il peut être préférable de ne pas le faire, à la fois pour une raison d'efficacité de l'estimation (la précision des valeurs estimées est meilleure sans pondération) et pour une raison pratique (utiliser les bonnes pondérations exige qu'on les recalcule sur la base des variables x). Enfin, si la liste x n'est pas complète non plus, alors quoi qu'on fasse les estimations seront biaisées.

En pratique, si on décide de pondérer les observations par la variable appelée, mettons, `poids`, les instructions sont les suivantes :

```
proc logistic data=tab descending covout outest=param;
  model secondeg = &listvar;
  weight poids/normalize;
run;
```

où `&listvar` est la liste des variables x. Noter l'option `normalize` de l'instruction `weight`, indispensable pour garantir des poids normalisés et, partant, des écarts-types corrects (voir *supra*).

Pour le calcul des effets marginaux, si les individus ne pèsent pas du même poids, alors il faut pondérer les observations. Pour ce faire, on renseigne le paramètre `ponder=` de la macro `marginal` par le nom de la variable de pondération. Par exemple :

```
%marginal(tab_ent=tab,x=&listvar,
          param_ent=param,var_qual=zep,ponder=poids);
```

II.9 En guise de conclusion : petit guide de conduite d'une étude

De manière très générale, la conduite d'une étude passe par (au moins) trois étapes :

- bien clarifier la finalité de l'étude et organiser les données en conséquence ;
- justifier autant que faire se peut l'utilisation du modèle *logit* pour traiter le problème ;
- présenter de manière la plus lisible possible, avec les outils adéquats, les résultats de l'analyse.

Explicitons ces trois points.

Première étape : clarifier la finalité de l'étude. Il s'agit d'abord de choisir entre les deux démarches offertes par la modélisation : (1) centrer l'analyse sur une variable principale, comme ce qui a été fait avec la variable `zep` (jusqu'à la section II.5) ; (2) ou bien se livrer à une analyse discriminante et identifier les variables les plus discriminantes (sections II.6 et II.7). Dans le premier cas, on hiérarchise *a priori* les variables en en distinguant une – la variable principale – sur laquelle on centre l'analyse et en conférant aux autres le statut de variables de contrôle. La finalité est de neutraliser les effets de *effets de structure* (ou *effets de composition*) qui faussent le lien entre la variable principale et la variable d'intérêt. Dans le second cas, on n'instaure pas de *distinguo a priori* entre les variables, mais l'analyse doit conduire, en règle très générale, à les hiérarchiser. Dans les deux cas de figure, les variables doivent être choisies et organisées avec le plus grand soin.

La première démarche exige une qualité quasi irréprochable de la variable principale, centrale dans l'analyse. On ne peut admettre, par exemple, de valeurs manquantes. Si le cas se présente, il faut se résoudre à supprimer les observations concernées, quitte à redresser l'échantillon résultant si nécessaire. En revanche, on peut être un peu moins regardant sur les variables de contrôle de par leur statut (relativement) secondaire. On peut s'accommoder de valeurs manquantes en les traitant en conséquence (section II.3.c). Autre point d'attention, le choix des variables de contrôle qui permettront de neutraliser au moins en partie les *effets de structure* doit être pesé. Le cas du milieu social de l'élève qu'on a introduit comme variable de contrôle dans le modèle *logit* (section II.3.c) en est une illustration. Le zonage de l'éducation prioritaire, défini au début des années 1980, reposait sur la catégorie sociale des élèves. En principe, les établissements scolarisant une proportion importante d'élèves de milieux sociaux défavorisés ont été affectés en éducation prioritaire. Dès lors, à partir du moment où la catégorie sociale apparait comme intimement liée à l'éducation prioritaire, comment justifier le fait de la décorréler de la dimension ZEP ? On peut s'autoriser à le faire en arguant que ce critère social n'a pas été strictement respecté dans la pratique, et ajouter que la catégorie sociale capte d'autres dimensions que la difficulté scolaire, qui est le cœur de cible de l'éducation prioritaire. Soit. On en reste alors au constat, établi en section II.3.d, du rôle positif de l'éducation priori-

taire sur le passage en seconde générale ou technologique. Peut-on aller plus loin, introduire d'autres variables de contrôle ? On sait que des moyens plus importants ont été affectés aux établissements relevant de l'éducation prioritaire. Les classes sont moins nombreuses qu'ailleurs. Ceci pourrait expliquer en partie cela : si on pense que des classes moins nombreuses favorisent les apprentissages et permettent aux élèves concernés d'être mieux préparés à la seconde générale, alors l'impact positif de l'appartenance à une ZEP en est peut-être la conséquence. Faut-il alors raisonner à taille de classe fixée, au risque de vider l'éducation prioritaire de toute substance et d'en faire une coquille vide ?

Si on choisit une démarche de type analyse discriminante où les variables ont le même statut, il faut d'abord s'assurer de la qualité de chacune d'elles. Les éventuelles valeurs manquantes doivent être traitées (voir section II.3.c). Ensuite, il faut bien choisir ses variables, il faut les organiser dans la perspective de répondre à la question : au bout du compte, parmi tous les facteurs qui distinguent les deux catégories d'individus, quels sont ceux qui jouent le plus grand rôle ? Surtout si elles sont nombreuses, il est utile de les regrouper en familles. Par exemple, et pour rester dans le domaine de l'éducation, si on dispose de variables sur la catégorie sociale des parents des élèves, sur leurs diplômes, sur le niveau de leurs revenus, on peut envisager de les mettre ensemble sous une rubrique « environnement familial de l'élève ». Ceci pourra faciliter les commentaires. Le cas échéant, on les sélectionnera pour éviter l'impression de « mélanger des choux et des carottes », ou pour la clarté des conclusions auxquelles on souhaite aboutir (voir à ce propos la discussion, page 89, sur la variable d'âge).

Quelle que soit la démarche employée, on s'attachera à bien définir la modalité de référence (voir la section I.2.c) et on veillera à la parcimonie du modèle, en évitant en particulier un nombre trop important de modalités pour les variables polytomiques (voir section I.4.a). On pourra utilement croiser la variable d'intérêt avec chaque variable (polytomique) du modèle. Cela permettra notamment de repérer les modalités rares (i.e. à effectif insuffisant), de les regrouper avec d'autres qui lui sont proches ou que l'on considère comme telles. C'est par ailleurs un bon moyen de prendre connaissance des données.

Deuxième étape, la justification du *logit*. Il s'agit de convaincre le lecteur de la nécessité d'utiliser un modèle statistique pour répondre aux questions posées. De « simples » tables présentant des statistiques descriptives suffisent amplement, à condition de bien les choisir.

Lorsque l'analyse est centrée sur une variable principale, on commence par la croiser avec la variable d'intérêt. Dans l'exemple qui a été traité, le croisement des variables `secondeg` et `zep` a conduit à comparer deux proportions, le pourcentage d'élèves de ZEP passant en seconde générale ($P(y = 1|zep = 1)$) et le pourcentage d'élèves hors ZEP passant en seconde générale ($P(y = 1|zep = 0)$) – voir table 1. Puis, pour justifier le fait qu'on ne peut s'arrêter à cette comparaison, il faut trouver

une variable de contrôle qui est corrélée à la fois à la variable principale et à la variable d'intérêt. Dans notre exemple, le choix s'est porté sur la variable d'âge à l'entrée en 6ème, ce qui a conduit aux tables 2 et 3. La variable de contrôle est responsable d'*effets de structure* (ou *effets de composition*), qui expliquent une partie de l'écart constaté au départ. Pour les neutraliser, c'est-à-dire créer une situation (fictive) où ils n'existeraient pas, il faut recourir à un modèle.

Lorsqu'on suit une démarche d'analyse discriminante, la justification est de même nature. On choisit deux variables, que l'on souhaite mettre en avant dans la démarche (la catégorie sociale et le niveau en 6ème dans notre exemple – voir section II.6), qui sont corrélées. L'apport de la modélisation est de les décorréler pour savoir laquelle joue le premier rôle.

Dernier point : la présentation des résultats. Là aussi, elle dépend de la démarche. Avec une variable principale, l'outil à privilégier est l'effet marginal de la variable, conditionnellement aux (compte tenu des) variables de contrôle retenues dans le modèle. Son calcul et sa comparaison à l'écart « brut » permettent, le cas échéant, de calculer la part de l'écart brut expliquée par les variables prises en compte. C'est par exemple ce qui a été fait avec la seule variable de catégorie sociale, qui explique environ 67% de l'écart brut constaté de 13,5 points entre les élèves de ZEP et les autres (page 76). Cela étant, le modèle complet, c'est-à-dire avec l'ensemble des variables y compris le niveau de l'élève à l'entrée en 6ème, fait davantage qu'expliquer l'écart dans l'orientation puisqu'il inverse le signe de l'impact de l'éducation prioritaire qui, de négatif, devient positif. Il faut bien voir que ce cas de figure arrive très rarement en pratique : en règle très générale, la prise en compte des variables de contrôle fait varier l'effet marginal de la variable principale mais ne change pas le signe. Dans notre cas très spécifique, on a deux solutions. La première consiste à énoncer les résultats du modèle en étant le plus rigoureux possible et en essayant de trouver la formulation la moins lourde possible (voir un exemple d'énoncé page 73 en toute fin de section II.4). La seconde solution est, comme suggéré en section II.5, de se passer de modèle ...

Dans le cas d'une analyse discriminante, on a intérêt à systématiser le calcul des effets marginaux et à les présenter en face des écarts bruts (comme ce qui a été fait pour la table 10 de la section II.6). On utilisera la gamme d'outils disponibles pour répondre autant que faire se peut à la question de la hiérarchie des facteurs discriminants (voir section II.7).

D'une manière générale, on se gardera d'employer des formulations ambiguës. On a souligné à plusieurs reprises le caractère inapproprié de l'expression toute faite « *toutes choses égales par ailleurs* ». On évitera aussi le qualificatif « *impact significatif* » lorsqu'on commente les seuils de significativité, car le lecteur pourrait traduire par « *impact important* », ce qui n'est pas la même chose.

En conclusion, le modèle *logit* peut beaucoup apporter à l'analyse à condition de l'utiliser à bon escient et de ne pas lui attribuer une ambition démesurée. No-

tamment, il ne faut pas faire croire qu'il permet de mesurer un effet *causal*. Les résultats restent conditionnels aux variables introduites. Son premier objectif est d'aller au-delà des apparences (en neutralisant les *effets de structure* dans le cas d'une analyse centrée sur une variable principale, en décorrélant les variables entre elles dans le cas d'une analyse discriminante) et, ce faisant, de produire des constats parfois inattendus, susceptibles d'orienter de nouvelles investigations.

– Annexe –

La macro SAS de calcul des effets marginaux

La macro SAS, nommée *marginal*, utilise la procédure *iml* proposée par SAS dans un module spécifique, qui permet de faire du calcul matriciel. La structure de la macro est la suivante :

```
%macro marginal(tab_ent=,x=,param_ent=,var_qual=,ponder=);
    /* etape prealable, executee si il y a une variable de pondération */
%if &ponder ne %then %do;
  proc summary data=&tab_ent(keep=&ponder);
    var &ponder;
    output out=poidsm(keep=poidsm) mean=poidsm;
  run;
  data &tab_ent(drop=poidsm);
    if _n_=1 then set poidsm;
    set &tab_ent;
    poids=&ponder/poidsm;
  run;
  proc delete data=poidsm;run;
%end;
    /* calcul et impression de l'effet marginal */
proc iml;
 start lecture;
   (...)
 finish lecture;
 start effet;
   (...)
 finish effet;
 start impress;
   (...)
 finish impress;
 run lecture;
 run effet;
 run impress;
quit;
%if &syserr ne 0 %then %do;
 data _message_;
  message="Attention ! Erreur !";
 run;
 proc print data=_message_ noobs;
  var message;
 run;
 proc delete data=_message_;run;
%end;
%mend;
```

La macro a cinq paramètres : `tab_ent` nomme la table contenant les données en entrée de l'analyse, `x` donne la liste de toutes les variables x du modèle, `parm_ent`

nomme la table des valeurs estimées des paramètres issue de la procédure `logistic`, `var_qual` liste les modalités de la variable (une seule dans le cas d'une variable binaire, $p-1$ dans le cas d'une variable polytomique à p modalités) dont on calcule l'effet marginal, `ponder` donne le nom de la variable de poids (si elle existe).
La macro débute par une étape préalable, qui est exécutée si les individus de l'échantillon ne pèsent pas du même poids, auquel cas la variable de poids doit être déclarée par le paramètre `ponder`). Cette étape permet de normaliser la pondération (i.e. faire en sorte que la somme des poids soit égale à l'effectif de l'échantillon).
La macro lance ensuite la procédure `iml`. Elle se compose de trois modules. Le premier, `lecture`, transforme les données conservées dans des tables SAS en matrices ou vecteurs. Le second module, `effet`, calcule les effets marginaux de la (ou des) variable(s) sélectionnée(s). Le dernier, `impress`, imprime les résultats des calculs. Ces trois modules sont successivement exécutés par la commande `run`.
Enfin, elle se termine par des instructions d'impression d'un message d'erreur en cas de problème.
On détaille maintenant le contenu de chacun des trois modules.

Le module *lecture*
Son contenu est le suivant :

```
start lecture;
 use &tab_ent;read all var{&x} into x;
 use &param_ent;read all var{intercept &x}
                    where (_type_='PARMS') into b;
 use &param_ent;read all var{intercept &x}
                    where (_type_='COV') into cov;
 %if &ponder ne %then %do;
   use &tab_ent;read all var{poids} into poids;
 %end;
 n=nrow(x);
 x=j(n,1,1)||x;
 beta=t(b);
 explic={&x};
 qual={&var_qual};
finish lecture;
```

La première déclaration `use &tab_ent ...` part de la table SAS des données individuelles. Toutes les observations (option `all`) sont lues mais seules les variables sélectionnées par la clause `var{}` sont conservées. Les observations et les variables sont « versées » dans une matrice nommée `x`. Chaque ligne de la matrice correspond à une observation de la table SAS en entrée, et le nombre de colonnes de `x` est égal au nombre de variables sélectionnées.
La deuxième déclaration, `use &param_ent ...`, part de la table SAS issue de la procédure `logistic`, qui contient les valeurs estimées des paramètres ainsi que leurs variances et covariances. Elle ne retient qu'une observation (clause `where`), celle qui,

dans la table SAS, correspond à `_type_='PARMS'` (i.e. les valeurs des paramètres). Ces valeurs sont conservées dans le vecteur-ligne nommé `b`.
La troisième déclaration `use` extrait – clause `where` – de la même table SAS les valeurs des variances et covariances des paramètres, et les range dans la matrice nommée `cov`. La matrice `cov` est ainsi une matrice carrée de dimension égale au nombre de variables introduites dans le modèle auxquelles on ajoute le terme contant (dont le paramètre associé s'appelle, par défaut, `intercept`).
Enfin, la quatrième est optionnelle, car elle dépend de l'existence d'une variable de pondération. Elle crée le vecteur à une seule colonne contenant le poids.
La fonction `nrow` retourne le nombre de lignes de la matrice, nombre représenté ici par `n`.
`j(n,1,1)` représente une matrice de dimension $n \times 1$ (`n` premier paramètre de `j`, 1 deuxième paramètre de `j`), dont les valeurs valent toutes 1 (troisième paramètre de `j`). En bref, il s'agit du vecteur colonne composé de 1. Le signe `||` signifie que l'on apparie ligne à ligne les matrices `j(n,1,1)` et `x`, pour en faire une nouvelle matrice, dont on a conservé le nom `x`. Ce faisant, on ajoute à la matrice `x` une colonne supplémentaire qui représente le « terme constant » du modèle.
Le vecteur des paramètres `beta` est le transposé de `b`. C'est donc un vecteur-colonne, conformément à sa représentation adoptée lors de la présentation formelle du modèle (page 9).
`explic` est le vecteur-ligne qui contient les noms des variables du modèle, `qual` est le vecteur-ligne qui contient le nom de la variable qualitative du modèle dont on veut calculer l'effet marginal. Noter que dans le cas d'une variable polytomique, le vecteur `qual` a plusieurs composantes.

Le module *effet*
Le contenu du module de calcul des effets marginaux est le suivant :

```
start effet;
   /* on repère le rang, dans la liste &x des variables du modele, de la
      variable qualitative &var_qual (ou de la 1ere variable de la liste
      &var_qual s'il s'agit d'une variable polytomique) */
 r=0;
 do q=1 to ncol(explic);
   if explic[q]=qual[1] then r=q;
 end;
   /*** initialisation des grandeurs utilisees ... */
     /* ... pour le calcul des effets marginaux */
 delta=j(n,ncol(qual),0);
 delta_moy=j(ncol(qual),1,0);
     /* ... pour le calcul des ecarts-types */
 gradi=j(1,ncol(x),0);
 grad=j(ncol(qual),ncol(x),0);
 sigma=j(ncol(qual),1,0);  *ecart-type de l'effet marginal;
 p_value=j(ncol(qual),1,0);*seuil de significativite de l'effet marginal;
   /*** calcul des effets marginaux */
```

```
    /* situation où var_qual=0 */
 x[,r+1:r+ncol(qual)]=j(n,ncol(qual),0);
 x_0=x;
 g0=1/(1+exp(-x_0*beta));
 %if &ponder ne %then %do;
   g0=g0#poids;
 %end;
    /* situation où var_qual=1 */
 do j=1 to ncol(qual);
   x=x_0;
   x[,r+j]=j(n,1,1);
   g=1/(1+exp(-x*beta));
   %if &ponder ne %then %do;
     g=g#poids;
   %end;
   delta[,j]=g-g0;
   delta_moy=t(delta[+,]*100/n);
     /* calcul de l'écart-type */
   do i=1 to n;
     gradi=x[i,]#(g[i]#(1-g[i])-g0[i]#(1-g0[i]));
     gradi[1,r+j]=g[i]#(1-g[i]);
     grad[j,]=grad[j,]+gradi[1,];
   end;
   grad[j,]=grad[j,]/n;
   sigma[j]=sqrt(grad[j,]*cov*t(grad[j,]))*100;
 end;
 p_value=2*(1-probnorm(abs(delta_moy)/sigma));
finish effet;
```

Le module commence par repérer, dans la liste des variables introduites dans le modèle par l'instruction `model` de la procédure `logistic`, la variable dont on veut calculer l'effet marginal. Si cette variable est une variable dichotomique (comme la variable `zep`) alors `ncol(qual)` – nombre de colonnes du vecteur `qual` (voir module `lecture`) – est égal à 1.
Le module calcule ensuite l'effet marginal, en appliquant la formule (27) dans les cas d'une seule variable binaire, ou les formules de type (29) dans le cas d'une variable polytomique. On notera que l'instruction SAS `1/(1+exp(-x*beta))`, par exemple, est l'exacte transcription de la formule $G = 1/[1 + e^{-x\beta}]$. Cette facilité d'écriture (le passage simple de l'expression formelle en instructions SAS) est rendue possible par la convention que nous avons établie page 9 sur les représentations des variables x en vecteur-ligne et des paramètres β en vecteur-colonne (voir la note 4 page 9).
Les résultats sont pondérés si une pondération existe.
La grandeur `delta_moy` donne l'effet marginal de la variable.
La boucle `do j=1 to ncol(qual)` est effective si `ncol(qual)` est supérieur à 1, c'est-à-dire si on a affaire à une variable polytomique.
Enfin, la partie du module consacrée au calcul de l'écart-type, est l'application de la *méthode delta* dans le cas où $\theta = h(\beta)$ est l'effet marginal Δ (voir (32)).

Le module *impress*

Le module *impress* s'écrit :

```
 start impress;
    /* impression des résultats */
  delta_moyc=char(delta_moy,10,4);
  sigmac=char(sigma,10,4);
  p_valuec=char(p_value,12,4);
  do j=1 to ncol(qual);
    if p_value[j]<0.0001 then p_valuec[j]="    < 0.0001";
  end;
  effetc=delta_moyc||sigmac||p_valuec;
  noms_ligne=rowcat(t({&var_qual})||j(ncol(qual),1,"   "));
  noms_col={"effet marginal","  écart_type","  significativité"};
  mattrib effetc  rowname=noms_ligne
                  colname=noms_col
                  label=" ";
  print "Effet marginal de '&var_qual'";
  print effetc;
 finish impress;
quit;
```

Il imprime trois grandeurs : l'effet marginal de la variable, son écart-type et son seuil de significativité.

Index

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS - K.MOUSSALAM