REVUE DE STATISTIQUE APPLIQUÉE

C. GELPÉROWIC

Partitionnement récursif et régression: comparaison dans le cas de la prévision de risque à partir des courbes de sélection

Revue de statistique appliquée, tome 48, n° 4 (2000), p. 5-28

<http://www.numdam.org/item?id=RSA_2000__48_4_5_0>

Rev. Statistique Appliquée, 2000, **XLVIII** (4), 5-28

PARTITIONNEMENT RÉCURSIF ET RÉGRESSION : COMPARAISON DANS LE CAS DE LA PRÉVISION DE RISQUE À PARTIR DES COURBES DE SÉLECTION

C. Gelpérowic

Ceremade, Université Paris-IX Dauphine, 75775 Paris cedex 16, France

RÉSUMÉ

Les courbes de sélection permettent de définir un ordre sur les fonctions de score construites généralement par analyse logit ou analyse discriminante pour faire de la sélection de clientèle en finance ou en assurance. Cet article étend l'utilisation de ces courbes aux scores obtenus par segmentation et décrit diverses façons de les utiliser dans ce cas.

Mots-clés : *scores, analyse logit et régression par arbre, courbes de sélection, élagage, biais de sélection.*

ABSTRACT

Selection curves define an ordering on the scoring functions used in credit or insurance problems and derived from logit or discriminant analysis. This article extend their use for scoring functions obtained from regression tree methodology and describe various practical uses in this case.

Keywords : *Scoring functions, logit analysis and regression trees, selection curves, pruning, sample selection bias*

1. Introduction

De nombreuses méthodes, regroupées sous la dénomination de *méthodes de discrimination* ou *méthodes d'analyse discriminante*, sont utilisées dans des secteurs variés de la vie économique : médecine, finance, assurance, marketing.... Les approches sont différentes et reflètent la culture statistique des utilisateurs (économètres, informaticiens, statisticiens...). Ces méthodes, qui s'appliquent à des populations munies d'une partition et décrites par un ensemble de variables explicatives, ont pour objectif de séparer au mieux (discriminer) les classes de la partition, dans un but descriptif ou pour faire des prévisions. Elles peuvent être utilisées notamment pour prévoir des «risques» et faire de la sélection (sélection de «clientèle» en finance, dans le domaine des assurances, en marketing, pour le

choix d'un traitement en médecine, etc...). Dans tous ces cas, la règle construite, également appelée score, fournit un classement des observations, d'où leur nom de *méthodes de scoring*.

La régression (logistique, probit), le partitionnement récursif (également appelé segmentation par arbre binaire) tout comme l'analyse discriminante (linéaire ou quadratique) et les réseaux de neuronnes apparaissent comme les méthodes de discrimination les plus employées. De nombreux articles comparent sur le plan théorique ou à partir d'échantillon de données particuliers quelques unes de ces méthodes. La régression et l'analyse discriminante linéaire ont fait l'objet du plus grand nombre d'études depuis le début des années soixante-dix ([6], [11], [14], ...). Cependant, l'amélioration des méthodes de partitionnement dûe notamment aux travaux de Breiman et coll [3] a entraîné un regain d'intérêt pour ces techniques qui ont permis de pallier quelques unes des lacunes reprochées à la régression ou à l'analyse discriminante linéaire. Les études de M. Bardos ([1], [2]) ou de H. Frydman et coll. [7] s'intéressent plus particulièrement à la comparaison entre partitionnement récursif et d'autres méthodes de discrimination dans le cas de la détection des défaillances d'entreprises.

Dans cet article, nous nous intéressons à l'évaluation et à la comparaison entre la régression (logistique/probit) et le partitionnement récursif lorsque ces deux méthodes sont employées pour faire de la sélection et prévoir des risques potentiels. L'algorithme de construction de l'arbre dans ce cas doit être adapté afin de fournir un score algébrique. La régression et le partitionnement conduisent alors à des règles de sélection assez semblables. On peut ainsi utiliser les courbes de sélection [8] pour évaluer la qualité de la règle obtenue par segmentation et la comparer au score logit/probit. Etendues aux règles construites à partir d'une méthode de partitionnement récursif, les courbes de sélection constituent un outil complémentaire intéressant pour l'évaluation de la règle et la comparaison avec d'autres règles. Contrairement au taux d'erreur de classement qui est le critère généralement employé dans ce cas ([1], [10]), ou au critère d'évaluation proposé par A. Ciampi et coll [5], les courbes permettent de comparer et d'évaluer des méthodes, quels que soient la répartition initiale de la population, la part de marché fixée et les coûts éventuels d'erreur de classement.

Calculées pour les règles obtenues par partitionnement récursif, les courbes de sélection peuvent aussi être utilisées pour évaluer les règles successives obtenues lors des différentes étapes de réduction du grand arbre (élagage et regroupement). Cela va permettre en particulier de choisir, à chaque étape, la réduction qui modifie le moins la qualité de la règle dans la plage des valeurs de part de marché souhaitée. Enfin, pour les applications auxquelles nous nous intéressons, l'échantillon disponible n'est pas toujours obtenu par tirage aléatoire dans la population d'origine mais a été soumis à une sélection préalable. Lorsque la règle de présélection n'est pas indépendante du groupe d'appartenance, cela entraine un biais au niveau de la construction de la règle; un tel biais est appelé *biais de sélection*. Les courbes de sélection pourront être utilisées pour mettre en évidence le biais de sélection sur la règle et aussi pour évaluer théoriquement l'amélioration d'une correction éventuelle (comparaison des courbes avant et après correction).

La première partie précise les notations et donne quelques exemples d'applications. La deuxième partie présente les deux méthodes dans le cas de la prévision de risque. Pour le score construit par partitionnement, l'algorithme de construction

utilisé par Frydman [7] pour les faillites d'entreprises ou par Breiman [3] pour les arbres de régression (étendus ici au cas où la variable à régresser est une variable dichotomique – indicatrice du groupe d'appartenance) est modifié de sorte que la règle obtenue par segmentation fournisse un score algébrique (arbre de prévision). La règle peut s'interpréter ainsi comme une régression non paramétrique, ce qui facilite la comparaison avec le score logit/probit et l'utilisation des courbes de sélection, qui sont présentées dans la troisième partie.

Enfin, dans une dernière partie, on illustre l'intérêt de ces courbes pour la règle obtenue par segmentation par différentes études de Monte Carlo : comparaison avec un score logistique ou probit, outil pour la réduction du grand arbre, mise en évidence du biais de sélection sur la règle.

2. Cadre et Notations

2.1. Cadre

On considère une population $\mathcal{P}$ partitionnée en deux groupes $\mathcal{P}_1$ et $\mathcal{P}_2$. Les éléments de $\mathcal{P}$, les *individus*, peuvent être des personnes physiques, des entreprises, des titres.... Selon le domaine, les deux sous-populations sont appelées population des individus défaillants et des individus non-défaillants, ou population des individus *risqués* et des individus *non-risqués*, etc.... Chaque individu de $\mathcal{P}$ est décrit par un ensemble X de p variables explicatives $X' = (x_1, x_2, ..., x_p)$, appelées également prédicteurs. Les composantes de X peuvent être de nature quelconque : continue ou discrète, et dans ce cas, quantitative ou qualitative.

L'appartenance à l'une des deux sous-populations est modélisée par une variable dichotomique Y. On notera pour la suite :

$$Y = \begin{cases} 0 & \text{si l'individu est défaillant, risqué, etc...} \\ 1 & \text{sinon.} \end{cases}$$

On s'intéresse alors à la probabilité (conditionnelle) de non-défaillance $P(Y = 1|X = x)$.

2.2. Objectif

A partir d'un N-échantillon des variables $(X_i, Y_i)_{1 \leqslant i \leqslant N}$, les deux méthodes conduisent à la construction d'une règle de sélection (un score), fonction des variables observables X que l'on note $S(X)$ dans la suite. Le score fournit un classement des observations par risque décroissant. On peut alors l'utiliser soit pour affecter toute observation $X = x$ à l'un des groupes exclusivement, soit pour prévoir la probabilité de non défaillance (fonction croissante du score). La régression donne, par construction, un score algébrique et la probabilité de non-défaillance associée. Pour la segmentation par arbre, la règle peut conduire à un score algébrique en choisissant une règle d'affectation adaptée pour les nœuds terminaux.

2.3. Exemple d'application

Ces méthodes sont utilisées à différentes fins :

– La détection des défaillances d'entreprises : les études de M. Bardos [1] ou de H. Frydman *et al.* [7] illustrent cette application, en étudiant les défaillances d'entreprises sur une période déterminée (3 ans, 6 ans, ...). Dans ce cas, le *risque* correspond à la défaillance possible d'une entreprise, pour un secteur d'activité particulier; les variables explicatives peuvent être des ratios financiers, le secteur d'activité, etc....

– La sélection de clientèle telle que le *credit-scoring*. Avant tout accord de crédit (à des particuliers ou à des entreprises), les établissements financiers doivent prévoir le *risque* de défaillance (non paiement d'une ou de plusieurs échéance(s) du demandeur) avant de prendre leur décision (accord ou non, et éventuellement avec des assurances complémentaires). Le *risque* peut également correspondre à un remboursement anticipé. Une telle sélection d'individus se pratique dans d'autres domaines tels que celui des assurances, pour le mailing, ou pour la sélection d'étudiants à l'entrée de l'université. Dans tous les cas, la règle est établie à partir de certaines caractéristiques observables des individus. Elle doit permettre de prévoir au mieux le risque individuel et en même temps elle fournit un classement des observations le plus fiable possible.

– Pour l'aide au diagnostic médical, les scores sont utilisés pour prévoir des risques et pour choisir un traitement. Le *risque* correspond à la présence d'un symptôme particulier, ou au type du symptôme (bénin/malin), au choix du traitement le mieux adapté, etc .... Là encore, les variables explicatives peuvent être quantitatives (mesure de taux...) ou qualitatives (présence/absence d'une caractéristique particulière).

2.4. Biais d'échantillonnage

Pour les exemples précédents, les échantillons à partir desquels les règles sont construites ne sont pas toujours obtenus par tirage aléatoire dans la population d'origine. Il est fréquent qu'ils aient été soumis à une sélection préalable : pour la détection des entreprises défaillantes, l'étude est effectuée pour un secteur d'activité donné; pour l'octroi de crédit, on ne peut examiner les défaillances que pour les individus auxquels le crédit a été accordé; pour les patients à partir desquels on compare l'efficacité de deux traitements, le traitement choisi n'est pas lié au hasard. A l'inverse, les individus eux-mêmes peuvent être à l'origine de cette sélection (autosélection).

Dans ces différents cas, si l'échantillon utilisé pour construire la règle a été soumis à une sélection préalable (par un tiers, ou par les individus eux-mêmes) il n'est pas représentatif de la population totale. Ce biais d'échantillonnage, appelé *biais de sélection* va avoir des conséquences sur la règle construite (biais des estimateurs pour les méthodes paramétriques, règle de sélection moins performante, etc...). Le score obtenu par chaque méthode (variables pertinentes, propriétés des estimateurs,...) et sa qualité à sélectionner vont ainsi être modifiés.

L'effet du biais de sélection et sa prise en compte pour «correction» va dépendre de chaque méthode. Dans le cas de la régression, le biais s'inscrit dans

le cadre des problèmes d'observabilités partielles étudiés par D.J. Poirier [15]. L'utilisation d'un modèle bivarié permet de prendre en compte la présélection au moment de la construction de la règle, et ainsi, d'en corriger partiellement les effets. En segmentation, le biais de sélection peut également apparaître (puisque les deux méthodes sont utilisées pour les mêmes applications); cependant il semble rarement évoqué lors de la construction puis l'utilisation de la règle. Les courbes de sélection vont permettre de mettre en évidence ce biais sur la règle, puis d'évaluer les effets des méthodes de correction proposées.

3. Régression et Partitionnement récursif

Cette section rappelle le principe des deux méthodes dans le cas de la prévision de risque.

3.1. La régression logistique et probit

La régression (voir par exemple G. Maddala [12]) permet de prévoir la probabilité de non-défaillance conditionnelle, par une approche paramétrique. Elle considère pour cela la loi conditionnelle de la variable $Y|X$ dont la loi $P(Y = 1|X = x)$ est de la forme $g(x;\theta)$; θ est un vecteur de paramètres à estimer ($\theta \in \mathbb{R}^p$) et g une fonction donnée. En choisissant pour $g(x;\theta)$ la fonction de répartition d'une loi logistique ou d'une loi normale, on obtient les modèles logit et probit respectivement. Cette représentation correspond à la régression linéaire de la variable latente Y^* sur les prédicteurs X :

$$Y^* = X'\beta + u$$

avec :

– u aléa centré, de variance σ^2 et de loi normale ou logistique, et de fonction de répartition F_u (symétrique).

– Y^* la variable latente liée à la variable observable Y par $Y = \mathbb{1}_{Y^* \geqslant 0}$.

La probabilité de non défaillance est ainsi une fonction croissante du score $S(X) = X'\beta : g(x;\theta) = F_u\left(x'\frac{\beta}{\sigma}\right)$. Seul le paramètre $\frac{\beta}{\sigma}$ est estimable, et on peut se ramener à $\sigma = 1$. Par la méthode du maximum de vraisemblance, on obtient un estimateur $\hat{\theta}$ du vecteur des paramètres θ. Pour chaque observation $X = x$, on calcule alors le score $x'\hat{\theta}$, et les observations sont classées par score croissant (correspondant à des probabilités de non défaillance estimées croissantes $\hat{g}(x) = g(x;\hat{\theta})$).

3.2. La segmentation par arbre binaire

La segmentation par arbre permet également d'obtenir une estimation de la probabilité de non défaillance, conditionnellement aux classes de la partition. En ce sens, elle peut être vue comme une régression non paramétrique de la variable latente Y^* sur les indicatrices des classes de la partition. Dans ce paragraphe, nous

utilisons un algorithme de construction proche de celui utilisé par Frydman [7] ou par Breiman [3] pour les arbres de régression (étendus au cas d'une variable latente). La règle obtenue fournit un classement des observations. Il va alors être possible de comparer la règle obtenue par segmentation à d'autres règles de discrimination, par comparaison des classements, à partir des courbes de sélection ([8]).

Ce critère d'évaluation et de comparaison est plus général que le taux d'erreur de classement utilisé habituellement. De plus, étendu aux arbres, il constitue un outil complémentaire pour l'élagage du grand arbre et la mise en évidence du biais de sélection.

3.2.1. Principe général de construction

Un arbre est construit par divisions dichotomiques successives d'un nœud a (sous-ensemble d'observations) en deux descendants, un nœud gauche a_g et un nœud droit a_d, en choisissant, à chaque étape (chaque nœud), la variable et la coupure sur cette variable qui réduit au plus le *mélange* des groupes dans les nœuds descendants (qui sont ainsi plus homogènes relativement à la variable de groupe Y).

Pour les différentes étapes de la construction de l'arbre (division, arrêt, élagage du grand arbre et décision) les critères employés peuvent varier en fonction des objectifs (classification, discrimination, prévision, etc...).

3.2.2. Construction de l'arbre

On précise dans ce paragraphe les critères que l'on utilise à chaque étape. On pourra consulter [9] pour une présentation plus détaillée de la méthode.

Division On utilise comme critère de division la réduction d'impureté qui, définie à partir de l'indice de Gini, et dans le cas de deux groupes correspond à la réduction de la variance interne d'un nœud. L'impureté, pour un nœud a, est définie par :

$$i(a) = P(Y = 1|X \in a)P(Y = 0|X \in a)$$

La division du nœud a en deux descendants ($a_G = \{x^* \leqslant c^*\}$ et $a_D = \{x^* > c^*\}$ dans le cas d'une variable x^* quantitative) est obtenue pour la variable x^* et la coupure c^* qui maximise la réduction d'impureté (donc la variance interne des nœuds) Δi définie par :

$$\Delta i = i(a) - (P(X \in a_G)i(a_G) + P(X \in a_D)i(a_D))$$

En itérant le procédé, on obtient ainsi un grand arbre.

Réduction du grand arbre Le grand arbre est ensuite «réduit» soit par retrait des branches les moins informatives (élagage), soit par regroupement des nœuds pour lesquels la décision est «proche» (dans le cas de la prévision, cela correspond à une différence des probabilités non significative), afin d'aboutir à un arbre «utilisable» constitué de J nœuds terminaux $A_1, A_2, ..., A_J$. La réduction du nombre de nœuds terminaux par élagage présente certains inconvénients. Si deux nœuds terminaux

N_1 et N_2 issus de branches différentes conduisent à des règles de décision assez «proches», il est préférable de regrouper ces deux nœuds plutôt que de procéder à un élagage de l'arbre qui peut faire disparaitre l'une des branches (*i.e.* l'un de ces deux nœuds). Cette méthode de «regroupement» a été initialement proposée par Ciampi ([4], [5]) pour la méthode SEGMAG (SEGMentation et AGrégation). Pour les arbres construits ici, on utilisera à la fois l'élagage et le regroupement pour réduire le grand arbre.

Décision On associe à chaque nœud terminal A_j, une décision $d_j(X)$. Dans le cas de la prévision, plutôt que d'affecter chaque feuille à l'un des groupes à l'exclusion de l'autre, on calcule la probabilité *a posteriori* d'appartenance au groupe $Y = 1$ conditionnellement à la feuille A_j : $p_j = P(Y = 1|X \in A_j)$. On note indifféremment $\hat{p}_j$ ou $\hat{d}_j$, l'estimation de cette probabilité, ou la décision associée. Cette décision correspond à la minimisation de l'erreur quadratique $E(Y - d(X))^2$ associée à la règle.

A partir d'un échantillon de taille N (autre que celui qui a servi à construire l'arbre), on obtient, pour chaque classe ou nœud A_j, des estimateurs $\hat{p}_j$ des probabilités d'appartenance au groupe $Y = 1$; plus précisèment, $\hat{p}_j$ est l'estimateur de la proportion d'individus non défaillants appartenant à la région associée à A_j. Ces estimateurs sont définis par :

$$\hat{p}_j = \frac{\sum_{i=1}^{N} Y_i \mathbb{1}_{X_i \in A_j}}{N_j}$$

où

– $\mathbb{1}_{X \in A_j}$ désigne l'indicatrice de la classe A_j ($\mathbb{1}_{X \in A_j} = 1$ si $X \in A_j$ et $\mathbb{1}_{X \in A_j} = 0$ sinon),

– et N_j l'effectif du nœud A_j.

Si l'échantillon est obtenu par tirage aléatoire dans $\mathcal{P}$, l'estimateur $\hat{p}_j$ est un estimateur convergent pour la probabilité conditionnelle $p_j = p(Y = 1|X \in A_j)$.

3.2.3. Utilisation de la segmentation pour la prévision de risque

Contrairement à la régression qui fournit un score algébrique, la règle obtenue à partir de l'arbre correspond à une partition de l'espace des variables explicatives, c'est-à-dire une suite d'assertions sur ces variables. On peut se ramener à un score algébrique en associant chaque feuille à sa décision p_j. On obtient ainsi une règle de classement des observations (par risque décroissant). On peut alors comparer les classements fournis par les deux méthodes à partir des courbes de sélection.

A partir des probabilités estimées $\hat{p}_j$, et après avoir réordonné les nœuds A_j par $\hat{p}_j$ croissants, on peut définir un score algébrique de la façon suivante :

$$\tilde{S}(X) = p_j \text{ pour } j = 1, ..., J \text{ et } X \in A_j$$

Contrairement à la régression, le classement obtenu à partir du score $\tilde{S}(X)$ n'est que partiel; il permet seulement de discriminer les observations entre J groupes (les J nœuds terminaux de l'arbre). Toutes les observations d'un même nœud sont considérées comme «*ex-aequo*». Pour ces observations, et si le nombre de nœuds J n'est pas assez élevé, il peut être utile, pour obtenir une discrimination plus *fine*, d'effectuer un traitement supplémentaire.

3.2.4. Régression non paramétrique particulière

La règle obtenue en construisant un arbre de prévision peut être considérée, *a posteriori*, comme une régression non paramétrique de la variable latente Y^* sur les indicatrices des classes de la partition (arbre de régression). Les variables explicatives (les indicatrices) sont sélectionnées au cours de la procédure de construction de l'arbre. La règle conduit à l'estimation, par des fonctions constantes par morceaux, de la fonction de probabilité :

$$P(Y = 1|X = x) = h(x) \tag{3.1}$$

où h, est une fonction inconnue, qui est approchée (estimée) sur A_j par une fonction constante : p_j.

En notant Y^* la variable latente liée à Y ($Y = \mathbb{1}_{Y^* \geqslant 0}$) cette probabilité correspond à la régression de Y^* sur une fonction des variables X :

$$Y^* = g(X) - \varepsilon \tag{3.2}$$

avec :

– X l'ensemble des variables explicatives,

– g une fonction inconnue,

– ε un aléa de fonction de répartition G_ε, et tel que $E(\varepsilon|X) = 0, V(\varepsilon|X) = \sigma^2$.

La probabilité de non défaillance est alors, d'après l'équation (3.2) :

$$P(Y = 1|X = x) = G_\varepsilon(g(x)) = h(x)$$

En effet, on a successivement : $P(Y = 1|X = x) = E(Y|X = x) = P(Y^* \geqslant 0|X = x) = P(\varepsilon \leqslant g(x)) = G_\varepsilon(g(x))$.

La segmentation conduit à une partition $A_1, A_2, ..., A_J$ telle que, pour toute valeur $x \in A_j$, $h(x)$ est approximée par $p_j(x)$. Ce qui peut s'interpréter, *a posteriori*, comme l'estimation (non paramétrique) de la fonction h par des fonctions «constantes» sur les sous-ensembles $A_j \in \chi$ (ces sous-ensembles sont des hypercubes de $\mathbb{R}^p$ si les variables explicatives sont continues).

En effet, en notant f la densité de X, on obtient l'expression de la probabilité de non-défaillance, conditionnellement à un nœud A_j, en utilisant l'équation (3.1) et

l'approximation p_j de $h(x)$ sur $x \in A_j$:

$$P(Y = 1|X \in A_j) = \frac{\int_{A_j} P(Y = 1|X = x) f(x) dx}{P(X \in A_j)}$$

$$\approx p_j \frac{\int_{A_j} f(x) dx}{P(X \in A_j)}$$

$$\approx p_j$$

4. Comparaison des deux méthodes

La régression logistique/probit et la segmentation par arbre, employées toutes les deux pour la prévision de risque, conduisent à des règles assez proches. Dans ce paragraphe, après une comparaison de l'expression des deux règles sur un exemple de «sélection de clientèle», nous allons voir comment les courbes de sélection [8] peuvent alors être utilisées pour évaluer et comparer ces règles entre elles.

4.1. Première comparaison arbre/régression (logistique ou probit)

Le partitionnement récursif et la régression considèrent tous deux la distribution conditionnelle de la variable de groupe $Y|X$.

La régression adopte une approche paramétrique pour estimer la probabilité de non défaillance conditionnelle $P(Y = 1|X = x) = g(x; \theta)$, tandis que la segmentation fournit une estimation de cette probabilité à partir d'une approche non paramétrique.

Les deux méthodes permettent de prendre en compte les interactions entre variables explicatives. Pour la règle obtenue par segmentation, les variables explicatives sont les indicatrices de la partition; elles sont déterminées au cours de la procédure de construction de la règle. Pour la régression logit /probit, on peut se ramener au cas où les régresseurs sont des indicatrices créées à partir des variables initiales. Il suffit pour cela de découper toute variable initiale continue en k classes, et de considérer les indicatrices associées (on retient $k - 1$ indicatrices afin que le modèle reste identifiable). La régression est alors effectuée sur les indicatrices des classes. Après une première régression sur ces variables simples, on rajoute parmi les variables explicatives des croisements entre deux puis trois variables, etc... Contrairement au cas de la règle obtenue par l'arbre, ces indicatrices ne forment pas une partition. De plus, elles sont déterminées *avant* l'estimation de la règle et non *pendant*.

Les deux méthodes permettent de traiter des échantillons de taille élevée et de considérer conjointement parmi les variables explicatives des variables discrètes et continues. Les courbes de sélection peuvent fournir un outil de comparaison supplémentaire.

4.2. Les courbes de sélection comme critère de comparaison

Un score S fournit un classement des observations et plusieurs modes de sélection, en fonction du seuil s choisi. La qualité d'un score, c'est-à-dire sa capacité à sélectionner les observations les plus fiables, pour lesquelles $P(Y = 1|X)$ est «grand», peut être évaluée quelle que soit la valeur du seuil, à partir de courbes [8] (courbes de sélection ou courbes de performance). Les règles de classement obtenues par la régression et la segmentation vont ainsi pouvoir être comparées à partir des courbes de sélection associées. Ces courbes trouvent diverses utilisations en segmentation : outre pour comparer la règle obtenue avec celle obtenue par régression, elles constituent un outil d'évaluation de la qualité de la règle, et d'aide pour l'étape de réduction du grand arbre. Elles peuvent aussi être employées pour la mise en évidence théorique des biais d'échantillonnage.

4.2.1. Les courbes de sélection

Elles définissent une relation d'ordre sur les scores : un score S_1 est préféré à un score S_2 si il conduit à une sélection plus précise des observations.

Définition 4.1. – *(Gouriéroux [8]) Pour un score S, les courbes de sélection sont des courbes paramétrées définies par :*

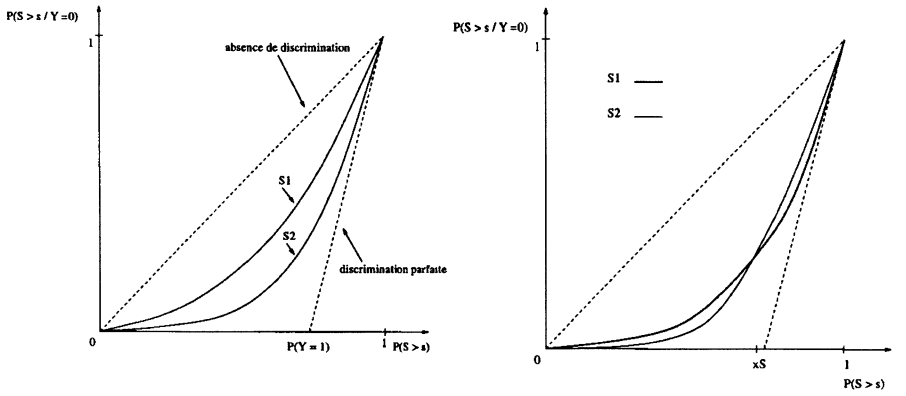
$$(Cs) \begin{cases} x(s) = P(S \geqslant s) \\ y(s) = P(S \geqslant s|Y = 0) \end{cases} \forall s \in \mathbb{R}$$

(i) La courbe est toujours située dans le pavé $[0;1] \times [0;1]$, elle passe par les points $(0;0)$ et $(1;1)$, et elle est invariante par transformation croissante du score. De plus, pour un score discriminant (tel que $x(s) > y(s)$), la courbe de sélection est croissante convexe, et toujours située en dessous de la première bissectrice.

(ii) Les courbes de sélection permettent d'évaluer la qualité d'un score, mais également de comparer différents scores entre eux. Ainsi, un score S_2 sera plus performant qu'un score S_1 (on notera $S_1 \prec S_2$) si la courbe de sélection associée au score S_2 est en dessous de celle associée au score S_1, comme l'illustre la figure 1-a. La courbe d'un score est, de plus, toujours comprise entre les courbes des deux scores «limites» : le score parfait (correspondant à $S = Y$) et un score non discriminant (dont la courbe a pour équation $y = x$).

(iii) Enfin, il peut arriver que les scores ne soient pas comparables globalement, c'est-à-dire quelle que soit la part de marché (abscisse de la courbe). On peut cependant faire la comparaison pour une part de marché donnée $x < x_S$ (figure 1-b), ou sur un intervalle $x_S \varepsilon [\underline{x}; \overline{x}]$ donné. Cette comparaison partielle est suffisante dans de nombreuses applications pour lesquelles le taux de sélection est fixé (mailing).

(iv) Si on considère l'hypothèse $H_0 : \{Y = 0\}$, contre $H_1 : \{Y = 1\}$ correspondant au test de l'hypothèse «l'observation X provient du groupe $Y = 0$» contre «l'observation X provient du groupe $Y = 1$». La comparaison de deux scores S_1 et S_2 à partir des courbes de sélection correspond à la comparaison de deux tests T_1 et T_2 définis à partir des statistiques de tests S_1 et S_2 pour l'hypothèse H_0 contre H_1. On préfère le test qui est le plus puissant sans biais.



a) b)

FIGURE 1
a) Courbes cas $S_1 \prec S_2$; b) Cas S_1 et S_2 non comparables globalement

4.2.2. Expression des courbes pour les deux règles

Dans le cas d'un score obtenu par régression, et d'un score obtenu par segmentation, il est possible de déterminer l'équation exacte des courbes de sélection, si on connaît la loi des variables explicatives X. Le score logistique ou probit construit à partir des indicatrices des classes et le score obtenu par segmentation ne sont pas continus; la courbe de sélection ne sera ni continue ni régulière. On la transforme par lissage.

4.2.3. Autres utilisations des courbes de sélection

Les courbes de sélection fournissent ainsi un outil d'évaluation et de comparaison entre la régression et la segmentation par arbre. Elles peuvent également être utilisées pour l'élagage du grand arbre ou pour mettre en évidence l'effet d'un biais de sélection.

Elagage En calculant à chaque étape de la réduction du grand arbre, la courbe de sélection associée, on peut mesurer – au niveau de la qualité de la règle – l'effet du retrait d'une branche ou du regroupement de plusieurs nœuds sur la règle, et ce, quelle que soit la répartition. La comparaison des courbes permet de choisir le regroupement ou l'élagage le plus satisfaisant. La dernière section illustre une telle utilisation. On peut noter que la comparaison des courbes (règles avant et après réduction) peut n'être que partielle, limitée à la plage de valeurs du taux de sélection correspondant à la part de marché fixée.

Evaluation d'une correction du biais de sélection Nous avons signalé que dans le cas de la prévision de risque, les échantillons à partir desquels les règles (logit/probit ou par arbre) sont construites, peuvent présenter un biais de sélection. Dans le cas de la régression, l'effet du biais sur la règle peut être mesuré par le biais qui en résulte au niveau des estimateurs. La prise en compte dans un modèle joint (présélection/défaillance) du processus de présélection et de sa corrélation avec le

processus de défaillance étudié permet alors de corriger partiellement ce biais (voir par exemple C. Meng et coll. [13]).

Pour les méthodes non paramétriques, en revanche, une telle démarche est difficilement envisageable, de par la nature de la méthode. Les courbes pourront cependant être utilisées pour mettre en évidence, par des simulations, l'effet d'un biais de sélection de l'échantillon sur la qualité de la règle. Pour cela, on comparera la règle obtenue sur l'échantillon complet et celle calculée sur l'échantillon disponible après présélection simulée. Il sera alors possible de comparer, pour différents biais de sélection simulés, l'effet de celui-ci respectivement sur la régression et la segmentation par arbre.

4.3. Exemple

On considère l'exemple de la sélection d'étudiants à l'entrée d'une université. La règle de sélection constitue un outil d'aide à la décision : en fournissant un classement ou une estimation de la prévision de la réussite future d'un étudiant, elle permet de ne sélectionner que ceux dont la probabilité de succès est la plus élevée, compte tenu d'éventuelles contraintes d'effectifs (part de marché). La réussite est représentée par la variable Y qui prend les modalités $Y = 1$ si l'étudiant réussit son année universitaire et $Y = 0$ sinon. Chaque observation est décrite par un ensemble X de variables explicatives, à partir desquelles on peut construire le score $S(X)$.

Les données On dispose des fichiers correspondants aux étudiants admis pour l'année 91-92 (Bac 91) et pour l'année 92-93 (Bac 92), qui se sont inscrits et pour lesquels la réussite ou l'échec en fin de première année de DEUG est observée. Les candidats peuvent être titulaires d'un Bac C ou d'un Bac D; par ailleurs, certains peuvent avoir redoublé leur classe Terminale. Les scores sur chacune de ces quatre sous-populations (Bac C/Bac D, redoublant/non redoublant) sont différents. On présente dans la suite le score construit pour les étudiants titulaires d'un Bac C et non redoublants, ce qui correspond à 234 observations.

Les variables retenue On suppose pour simplifier que cette sélection s'effectue à partir de trois variables explicatives disponibles au moment de la prise de décision, que l'on note X_1, X_2 et X_3. Ces variables correspondent respectivement à la note en mathématiques, la moyenne générale, et la qualité du lycée dans lequel l'étudiant effectue son année Terminale. X_1 et X_2 sont des variables continues (comprises entre O et 20), la variable X_3 est une variable discrète à 4 modalités codées de 1 à 4. Les deux premières variables sont continues (quantitatives), la troisième est discrète et ordonnée (qualitative ordonnée).

Régression Chaque variable continue (X_1 et X_2) est transformée en deux variables dichotomiques (trois classes), à partir de l'examen de la fonction de répartition empirique. Ainsi, pour X_1, on définit :

$$\mathbb{1}_1 = \mathbb{1}_{8 \leqslant X_1 \leqslant 10.5} \qquad \mathbb{1}_2 = \mathbb{1}_{X_1 > 10.5}$$

et pour X_2 :

$$\mathbb{1}_3 = \mathbb{1}_{9 < X_2 \leqslant 11} \qquad \mathbb{1}_4 = \mathbb{1}_{X_2 > 11}$$

La variable X_3, qui est discrète, est transformée quant à elle, en trois indicatrices, correspondant aux trois dernières modalités. On note $\mathbb{1}_5$, $\mathbb{1}_6$ et $\mathbb{1}_7$ ces trois indicatrices. On régresse Y^* (variable latente liée à Y) sur les sept variables ainsi créées et sur la constante.

$$Y^* = a_0 + \sum_{i=1}^{7} a_i \mathbb{1}_i + u$$

Après une première estimation qui détermine les variables significatives, on rajoute parmi les variables explicatives les croisements entre deux variables, puis trois, etc....

Segmentation Pour la segmentation, en revanche, aucune transformation préliminaire des variables n'est nécessaire. Les coupures sur chaque variable et les croisements entre variables sont obtenus au fur et à mesure de la construction de l'arbre.

Règles obtenues L'application des deux méthodes conduit aux deux règles suivantes qui peuvent être écrites soit sous forme d'arbre, soit sous forme d'une probabilité conditionnelle estimée.

(i) La régression pour le modèle probit fait ainsi apparaître trois variables explicatives :

– $\mathbb{1}_2$, l'indicatrice de la classe $X_1 > 10.5$ de la note de mathématiques :

– le croisement entre les classes des variables X_1 et X_2 : $\mathbb{1}_1 \times \mathbb{1}_4$;

– le croisement $\mathbb{1}_1 \times (\mathbb{1}_6 + \mathbb{1}_7)$.

Cela correspond à l'équation de régression :

$$P(Y = 1|X = x) = \Phi\{a_0 + a_1\mathbb{1}_2 + a_2\mathbb{1}_1 \times \mathbb{1}_4 + a_3\mathbb{1}_1 \times (\mathbb{1}_6 + \mathbb{1}_7)\} \quad (4.1)$$

où $\Phi(.)$ désigne la fonction de répartition de la loi normale centrée réduite.

La règle obtenue peut être représentée sous forme d'un arbre binaire composé de cinq nœuds, comme l'illustre la figure 2 page suivante. Des regroupements peuvent être faits entre certains nœuds terminaux (ce qui correspond au terme constant dans l'équation de régression).

Les décisions associées aux nœuds terminaux sont, dans le cas de la régression probit, et en notant d_j la décision associée au nœud N_j (pour $j = 1$ à 5) :

$$\begin{aligned} d_1 &= \Phi(a_0 + a_1) \quad d_2 = \Phi(a_0 + a_2 + a_3) \\ d_3 &= \Phi(a_0 + a_2) \quad d_4 = \Phi(a_0 + a_3) \\ \text{et} \quad d_5 &= \Phi(a_0). \end{aligned}$$

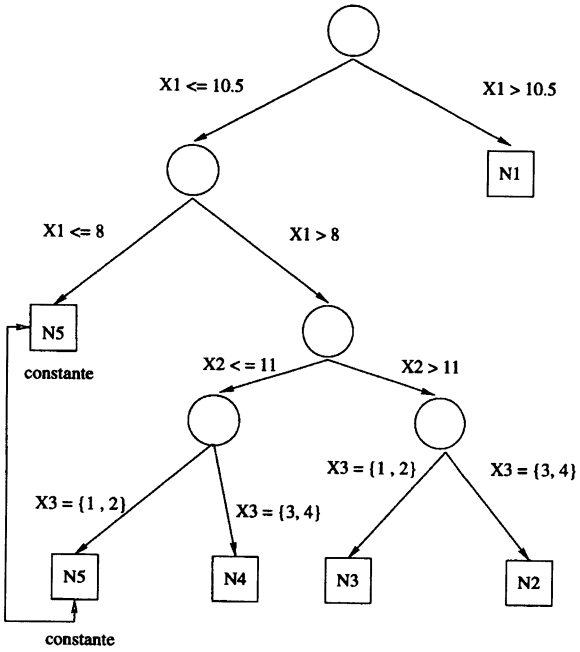


FIGURE 2
Représentation sous forme d'arbre de la règle obtenue par régression probit

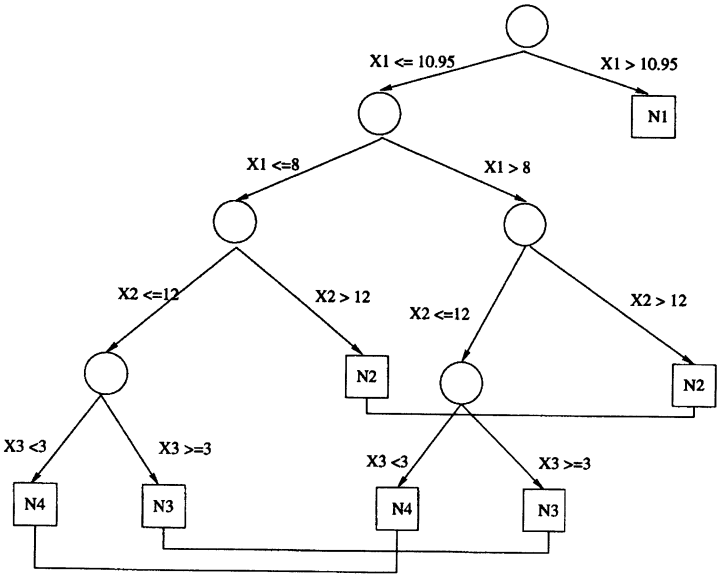


FIGURE 3
Règle obtenue par segmentation

(ii) La segmentation (*cf.* figure 3 page précédente) conduit à un arbre composé, après regroupements, de quatre nœuds terminaux N_1, N_2, N_3 et N_4, ce qui correspond a une partition de l'espace des variables explicatives en quatre classes :

– la première classe (nœud N_1) correspond aux observations pour lesquelles $\{X_1 > 10.95\}$;

– la deuxième classe de la partition (nœud N_2) correspond aux observations pour lesquelles $\{X_1 \leqslant 10.95\} \cap \{X_2 \geqslant 12\}$;

– la troisième classe est composée des observations pour lesquelles $\{X_1 \leqslant 10.95\} \cap \{X_2 \leqslant 12\} \cap \{X_3 \in \{3,4\}\}$;

– enfin, la quatrième classe (nœud N_4) correspond à $\{X_1 \leqslant 10.95\} \cap \{X_2 \leqslant 12\} \cap \{X_3 \in \{1,2\}\}$

La règle associée peut également s'écrire sous la forme d'une régression.

On a dans ce cas :

$$P(Y = 1|X = x) = p_1 \mathbb{1}_{\{x_1 > 10.95\}} + p_2 \mathbb{1}_{\{x_1 \leqslant 10.95\} \cap \{x_2 > 12\}}$$
$$+ p_3 \mathbb{1}_{\{X_1 < 10.95\} \cap \{X_2 \leqslant 12\} \cap \{X_3 \leqslant 3\}}$$
$$+ p_4 \mathbb{1}_{\{X_1 \leqslant 10.95\} \cap \{X_2 \leqslant 12\} \cap \{X_3 \leqslant 2\}}$$

Les deux règles conduisent à une partition assez proche pour certains nœuds. Les coupures observées pour les différentes variables sont parfois très voisines. Les mêmes croisements entre variables apparaissent dans les deux règles. La principale différence réside dans la possibilité, en régression paramétrique, d'avoir à la fois les indicatrices des variables initiales et les indicatrices des croisements dans l'expression de la probabilité, ce qui permet une modélisation plus fine.

La règle obtenue par segmentation pourrait être «affinée» en effectuant les coupures non plus sur'les variables simples, comme c'est le cas ici, mais sur une combinaison entre variables [5].

Ces fortes ressemblances sont confirmées par l'examen des courbes de sé lection qui sont pratiquement confondues.

5. Applications

5.1. Comparaison score probit et score obtenu par partitionnement récursif

On compare dans ce paragraphe les règles obtenues par régression et segmentation pour différents modèles simulés. Pour la régression, on choisit le modèle probit. Pour la segmentation, la division est effectuée sur une seule variable à la fois. La réduction du grand arbre est obtenue par la méthode décrite dans les paragraphes précédents (élagage/regroupement).

Tirage des échantillons On simule $n = 100$ échantillons ε de taille $N = 500$ à partir desquels on calcule la règle pour chaque méthode. Pour la règle obtenue par

segmentation, chaque échantillon est divisé en deux parties; la première partie fournit le partitionnement, la seconde est utilisée pour l'estimation des décisions associées aux nœuds terminaux. Les partitions obtenues sur les différents échantillons sont très voisines, tant au niveau des classes obtenues (variables discriminantes et coupures associées) que de leur nombre. Ceci s'explique en partie parce que les variables explicatives sont discrètes; ainsi, les valeurs possibles des coupures sont en nombre limité, et les mêmes coupures apparaissent lors des différentes simulations.

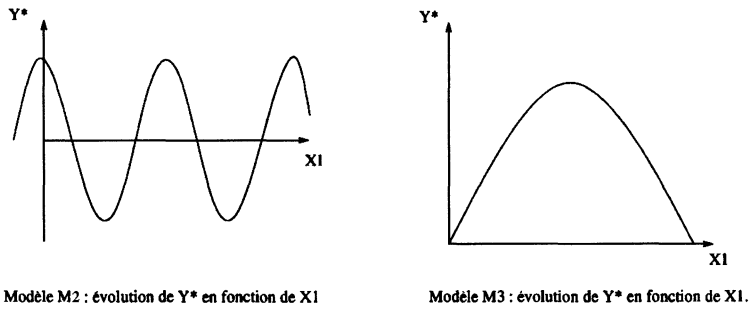
On fixe donc une partition de référence, par exemple la première partition obtenue lors des simulations; on note $\mathcal{A}$ cette partition qui est composée de J classes : $A_1, A_2, ..., A_J$. La $k^{\text{ième}}$ simulation fournit la partition $\mathcal{C}_k$. On associe alors à chaque classe de $\mathcal{C}_k$ la classe A_j de $\mathcal{A}$ qui lui est la plus liée. On note $C_{j,k}$ cette classe, ce qui revient à numéroter les classes de la $k^{\text{ième}}$ partition de la même façon que celles de la partition de référence. Soit alors $d_{j,k}$ la décision associée au nœud correspondant à la classe $C_{j,k}$. En faisant la moyenne sur les n échantillons, on obtient les décisions moyennes pour chaque nœud j ($\overline{d}_j = \frac{1}{n}\sum_{k=1}^{n} d_{j,k}$). Pour la régression, chaque échantillon fournit une estimation des coefficients du score; les n échantillons donnent ensuite les coefficients moyens de la régression. On a ainsi deux règles moyennes $S_1(X)$ et $S_2(X)$ (pour le score probit et le score obtenu par partitionnement respectivement). On construit alors les courbes de sélection moyenne $\overline{C}_1(X), \overline{C}_2(X)$ associées aux deux règles en faisant la moyenne des courbes obtenues pour $T = 100$ échantillons ε.

Modèles Simulés On considère trois modèles de la forme $Y^* = g(X;\beta) + u$ pour différentes expressions de la fonction g. La variable Y^* est une variable latente liée à la variable de groupe par $Y = \mathbb{1}_{Y^* \geqslant 0}$. Le vecteur X des variables explicatives est composé de deux variables X_1 et X_2. En segmentation, la variable ayant le plus grand nombre de modalités est souvent sélectionnée pour la meilleure division, afin d'éviter ce biais pour les comparaisons, on suppose que les variables explicatives ont le même nombre de modalités; de plus, elles suivent la même loi et peuvent éventuellement être corrélées. Leur coefficient de corrélation ρ prend les valeurs $\rho = 0$ ou 0.5.

Les modèles sont choisis pour prendre en compte différentes caractéristiques statistiques des données : interaction entre les variables explicatives, non linéarité du modèle, croissance monotone ou non du critère en fonction des variables explicatives, répartition initiale de la population, etc...

Pour chaque modèle, on fait varier la constante a pour avoir dans l'échantillon une proportion $P(Y = 1)$ égale respectivement à 0.25, 0.50 et 0.75. Dans M_1, on choisit les coefficients du modèle de sorte que X_1 et X_2 aient le même poids ($b = c = 1$). Dans les modèles M_2 et M_3, le coefficient du terme «non linéaire» a un poids plus important que le coefficient du terme linéaire. La distinction entre les modèles M_2 et M_3 réside dans la fonction liant X_1 et Y. Dans le deuxième cas (modèle M_3), les valeurs extrêmes de X_1 correspondent à des valeurs faibles de $P(Y = 1)$ tandis que les valeurs intermédiaire de X_1 correspondent à une valeur de $P(Y = 1)$ plus élevée. Dans le modèle M_2, $P(Y = 1)$ est alternativement fonction croissante et décroissante de X_1 (voir figure 4).

	Modèles	Paramètres
M_1	$Y^* = a + bX_1 + cX_2 + dX_1X_2 + u$	$u \rightsquigarrow N(0;1)$ avec $X_1, X_2 \rightsquigarrow U\{0,1,...4\}$ $\rho = 0$, puis $\rho = 0.5$ $d = 0.5\,c = 0.5\,b = 0.5$
M_2	$Y^* = a + b * \cos(\pi * X_1) + c * X_2 + u$	$u \rightsquigarrow N(0;1)$ avec $X_1, X_2 \rightsquigarrow U\{0,1,...4\}$ $\rho = 0$, puis $\rho = 0.5$ $b = 2,\ c = 1$
M_3	$Y^* = a + b * \sin(\pi * X_1/4) + c * X_2 + u$	$u \rightsquigarrow N(0;1)$ avec $X_1, X_2 \rightsquigarrow U\{0,1,...4\}$ $\rho = 0$, puis $\rho = 0.5$ $b = 4,\ c = 1$



Modèle M2 : évolution de Y* en fonction de X1 Modèle M3 : évolution de Y* en fonction de X1.

FIGURE 4
Evolution de Y^* en fonction de X_1

Pour chaque modèle $M_1 - M_3$, on considère successivement le cas où X_1 et X_2 sont non ou «moyennement» corrélées ($\rho = 0.5$).

Résultats On présente ici quelques-unes des courbes obtenues pour les différents modèles simulés, pour différentes valeurs de la corrélation entre variables explicatives, mais pour une population équirépartie ($P(Y = 1) = 0.5$).

– **Modèle** M_1 Dans ce cas, et quelle que soit la répartition initiale, la régression conduit à une meilleure sélection. La différence n'est parfois pas très grande, du fait de l'importance du nombre des nœuds terminaux. Si les résultats sont proches, ceci est obtenu, pour l'arbre, au prix d'une plus grande complexité (huit nœuds terminaux pour quatre coefficients à estimer). Il est évident que plus le nombre des nœuds augmente, plus on se rapproche d'une très bonne sélection. A la «limite», on estime $P(Y = 1|X_1 = x_1, X_2 = x_2)$ point par point.

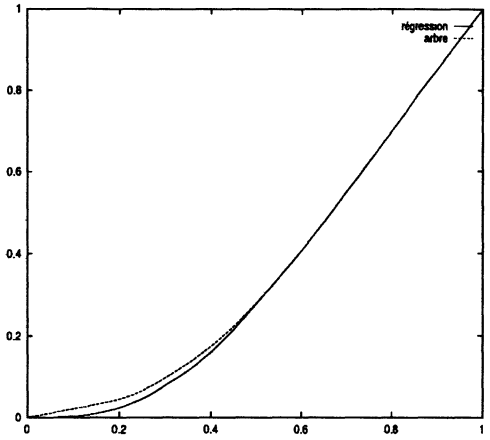


FIGURE 5
M_1 – *Cas* $P(Y = 1) = 0.5$ *et* $\rho = 0$

La figure 5 représente la courbe obtenue pour $\rho = 0$ et un arbre à 6 nœuds terminaux.

– **Modèle** M_2 Dans le cas où il y a non linéarité par rapport à l'une des deux variables explicatives (ici X_1), il n'y a pas de méthode qui domine de façon uniforme : les courbes se croisent. La règle obtenue par partitionnement conduit à un meilleur classement pour une sélection très sévère (abscisse de la courbe proche de 0 et inférieure au taux de non défaillance dans la population totale). En revanche le score obtenu par régression (toujours dans le cas d'équirépartition) domine pour une part de marché supérieure à celle de la population initiale ($x \geqslant 0.5$).

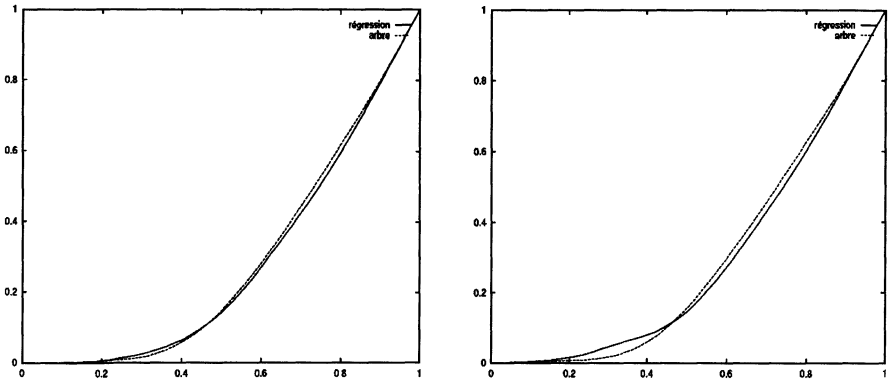


FIGURE 6
M_2 – $P(Y = 1) = 0.5$, *cas* $\rho = 0$ *et* $\rho = 0.5$

– **Modèle** M_3 Dans ce cas, la régression ne l'emporte jamais sur le partitionnement récursif. La différence entre les deux méthodes est d'autant plus nette que la probabilité de non défaillance ($P(Y = 1)$) augmente. Dans le cas où il y a corrélation entre les variables explicatives, la comparaison reste inchangée. On présente dans la figure 7 les courbes obtenues pour $\rho = 0$.

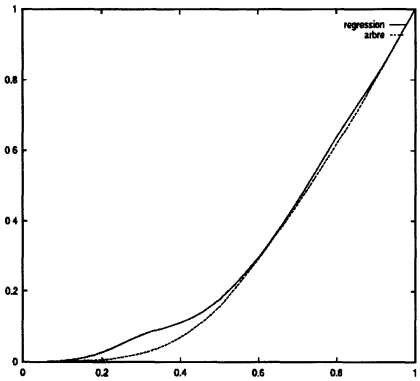


FIGURE 7
M_3 – $P(Y = 1) = 0.5, \rho = 0$

5.2. Utilisation des courbes pour la réduction du grand arbre

On considère le modèle M présenté dans le tableau ci-dessous. Ce modèle est dérivé du modèle proposé par Breiman [3] p. 238 pour les arbres de régression. On applique à M une segmentation à partir de l'algorithme décrit dans le paragraphe précédent. On calcule à chaque étape les courbes associées à la réduction du grand arbre (élagage, regroupement).

	Modèle	Paramètres
M	$\begin{cases} Y^* = -2 + X_2 + u \text{ si } X_1 = 0 \\ Y^* = -3 + 2X_2 + u \text{ si } X_1 = 1 \end{cases}$	où : $\begin{cases} u \rightsquigarrow N(0;1) \\ X_1 \rightsquigarrow \text{Bernouilli}\,(0.5) \\ X_2 \rightsquigarrow N(1.5;1) \end{cases}$

Les coefficients du modèle sont choisis de sorte que $P(Y = 1) = 0.4$. Comme pour l'exemple précédent, l'échantillon est de taille $N = 500$. La variable X_2 est transformée (tronquée) en une variable discrète à 5 modalités.

Courbes associées Le grand arbre est composé au maximum de 10 nœuds terminaux (2×5 modalités des deux variables). La première étape de la réduction ne modifie

pas beaucoup les règles (les courbes associées sont presque confondues). A la deuxième étape, on remarque une détérioration de la règle, seulement pour les scores élevés (valeur faible de l'abscisse de la courbe de sélection), puisque cette étape correspond à l'élagage d'une branche pour laquelle les observations $\{Y = 1\}$ sont fortement majoritaires (figure 8). Une plus nette détérioration apparaît à la quatrième étape du regroupement, comme l'illustre la figure 10 page suivante. En fonction de l'application, et de la part de marché fixée dans le cas où la règle est utilisée pour faire de la sélection, ces courbes peuvent aider dans le choix de l'arbre final.

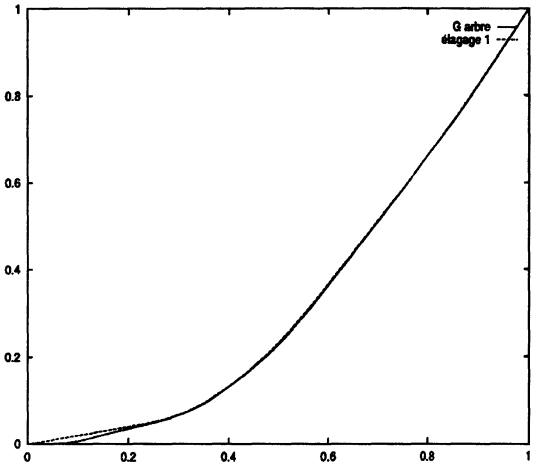


FIGURE 8
Courbes 2$^{\text{ième}}$ *étape*

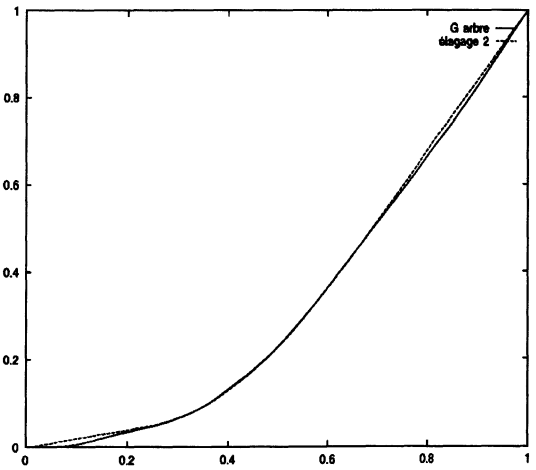


FIGURE 9
Courbe 3$^{\text{ième}}$ *étape*

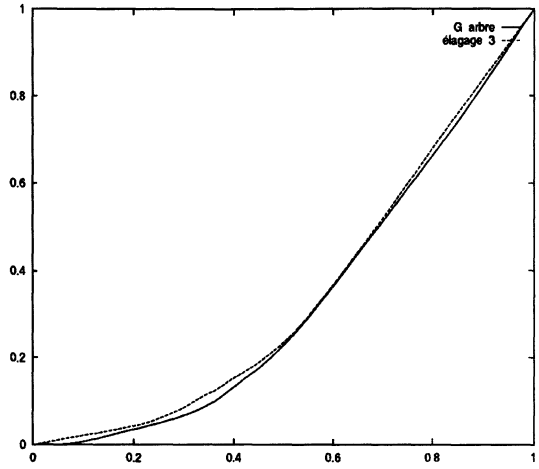


FIGURE 10
Courbes 4$^{\text{ième}}$ *étape*

Si, par exemple, on peut sélectionner 60 % de la population classée ($x = 0.6$ sur la courbe de sélection), on peut choisir l'arbre obtenu à la quatrième étape de la réduction (figure 10) puisque pour cette abscisse, la courbe associée est presque confondue avec celle du grand arbre. En revanche, pour une part de marché comprise entre 20 % et 50 % (supérieure à 75 %), il est préférable de choisir comme arbre de décision l'arbre obtenu à la deuxième ou à la troisième étape du regroupement (à la deuxième étape respectivement) puisque cet arbre discrimine mieux les observations dont le risque est plus faible (respectivement plus élevé).

5.3. Utilisation des courbes pour mettre en évidence le biais de sélection

Les courbes de sélection peuvent également être utilisées pour mettre en évidence le biais de sélection sur la règle. Si pour la régression logit/probit, la mise en évidence du biais se traduit par un biais au niveau des estimateurs, et ainsi du score estimé (coefficients significatifs), l'effet du biais sur la règle obtenue par partitionnement est plus difficile à mesurer. Ce biais peut avoir une incidence à la fois sur la partition obtenue (le grand arbre), sur le regroupement et sur les décisions associées aux nœuds terminaux retenus. Il est difficile de quantifier le biais exact pour ces trois étapes en dehors d'un modèle particulier. L'objet ici est simplement de montrer l'intérêt des courbes pour mesurer l'effet du biais sur la règle. Nous reprenons le modèle M simulé dans le paragraphe précédent (avec une légère modification de la valeur des coefficients). Pour obtenir un échantillon avec présélection, nous considérons le modèle bivarié ci-dessous. La première équation correspond au processus de présélection. La seconde équation décrit le modèle M. La corrélation entre les deux phénomènes est mesurée par la corrélation, notée ρ des résidus (u_1 et u_2) des deux équations.

$$Y_1^* = -1.5 + X_2 + u_1$$

$$\begin{cases} Y_2^* = -2 + X_2 + u_2 & \text{si} \quad X_1 = 0 \\ Y_2^* = -2.75 + 2X_2 + u_2 & \text{si} \quad X_1 = 1 \end{cases}$$

où Y_1^* et Y_2^* sont des variables latentes liées aux variables observables Y_1 et Y_2 par $Y_j = \mathbb{1}_{Y_j^* \geqslant 0}$. On simule un échantillon de taille $N = 500$ des variables (Y_1, Y_2, X_1, X_2), successivement pour $\rho = 0.25$, 0.50,0.75. On compare les courbes associées aux règles construites par partitionnement et obtenues pour le modèle M sur l'échantillon complet $(Y_{2,i}, X_{1,i}, X_{2,i})$ pour $1 \leqslant i \leqslant N$, et sur l'échantillon présélectionné $(Y_{2,i}, X_{1,i}, X_{2,i})$ pour $i = \{1 \leqslant i \leqslant N | Y_{1,i} = 1\}$.

Pour une faible corrélation, $\rho = 0.25$, les courbes obtenues à partir des deux échantillons ne sont pas très différentes. En revanche pour des valeurs de ρ plus élevées, la règle obtenue sur l'échantillon présélectionné est moins performante que celle construite à partir de l'échantillon complet, comme l'illustre la figure 11 ci-contre. On peut signaler que le grand arbre construit à partir de ce modèle conduit à la même partition quel que soit l'échantillon (présélectionné ou non). Pour cet exemple, c'est essentiellement au niveau de l'élagage et de l'estimation des probabilités des nœuds terminaux (quand ceux-ci correspondent à la même partition) que les différences apparaissent.

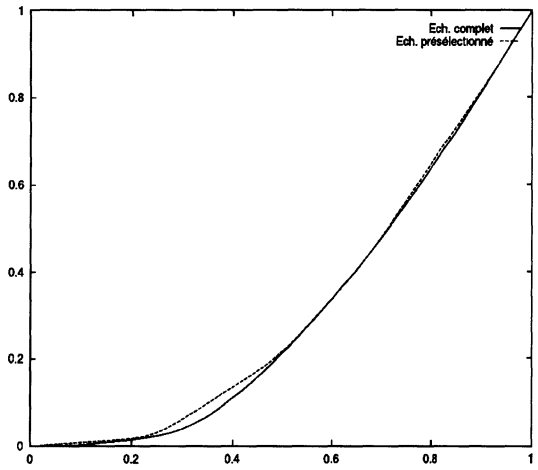


FIGURE 11
M – Courbes correspondant au partitionnement sur données complètes et présélectionnées – cas $\rho = 0.5$

Conclusion

Cet article utilise les courbes de sélection pour comparer les scores obtenus par partitionnement récursif et par régression (logit/probit). La construction des courbes de sélection associées au score obtenu par partitionnement récursif permet d'évaluer la qualité de la règle obtenue et constitue un bon outil de comparaison. Sur les

quelques modèles simulés et dans le cas où l'échantillon est obtenu par tirage aléatoire dans la population d'origine, il apparait qu'aucune méthode ne domine vraiment l'autre; d'autres modèles simulés pourraient conduire à d'autres conclusions. Dans les applications pratiques, l'utilisation conjointe des deux méthodes peut contribuer à obtenir une règle plus précise. Le partitionnement fournit les coupures discriminantes pour chaque variable, ainsi que les croisements entre variables, déterminant ainsi les indicatrices des classes sur lesquelles on peut alors faire la régression. Les courbes de sélection constituent de plus un outil d'évaluation utile pour les règles obtenues par segmentation.

Remerciements

L'auteur tient à remercier le Professeur Pierre Cazes dont les nombreuses remarques ont permis d'améliorer considérablement ce travail; ainsi que Madame Jacqueline Pradel qui est à l'origine de cette étude.

Références

[1] BARDOS M., *Trois méthodes d'analyse discriminante : Comparaison des résultats et confirmation de la qualité du score B pour les PME du batiment gros œuvre et du génie civil*, Cahiers économiques et monétaires, **33**, 19 89, 151-187.

[2] BARDOS M., ZHU W., *Comparaison de l'analyse discriminante linéaire et des réseaux de neurones. Application à la détection des défaillance d'entreprises*, Revue de Statistiques Appliquées, Vol. 45 No 4, 1997, 65-92.

[3] BREIMAN L., FRIEDMAN J.H., OLSHEN R.A., STONE C.J., *Classification and regression Trees*, Wadsworth International Group, 1984.

[4] CIAMPI A., CHANG C.H., HOGG S., MCKINNEY S., *Recursive partition : a versatile method for exploratory data analysis in biostatistics*, Biostatistics, 1987, 23-50.

[5] CIAMPI A., THIFFAULT J., SAGMAN U., *Evaluation de classification par le critère d'Akaike et la validation croisée*, Revue de Statistique appliquée, Vol. 36 No 3, 1988, 33-50.

[6] EFRON B., *The efflciency of logistic regression compared to norrnal dis criminant analysis*, Journal of the American Statistical Association, **70**, 1975, 892-898.

[7] FRYDMAN H., ALTMAN E. et KAO D.L., *Introducing recursive partitioning for financial classification : The case of financial distress*, Journal of Finance, **40**, 1985, 269-290.

[8] GOURIÉROUX C., *Courbes de performance, de sélection et de discrimination*, Annales d'économie et de statistique, Vol. 28, Oct.-Déc. 1992, 107-123.

[9] GUEGUEN A., NAKACHE J.P., *Méthode de discrimination basée sur la construction d'un arbre binaire* – Revue de statistique appliquée, Vol. 36 No 3, 1988, 19-38.

[10] HASTIE T., TIBSHIRANI R., BUJA A., *Flexible Discriminant Analysis by optimal scoring*, Journal of the American Statistical Association, **Vol. 89**, **No 428**, 1994, 1255-1270.

[11] MC FADDEN D., *A comment on discrimination analysis versw logit analysis*, Annals of Economic and Social Measurement, 1975, 511-523.

[12] MADDALA G.S., *Limited-dependent and qualitative variables*, Cambridge University Press, 1983.

[13] MENG C.L., SCHMIDT P., *On the cost of partial observability in the bivariate probit model*, International Economic Review, **26**, 1985, 71-85.

[14] PRESS S.J., WILSON S., *Choosing between logistic regression and discriminant analysis*, Journal of the American Statistical Association, **Vol. 73**, **No 364**, 1975, 699-705.

[15] POIRIER D.J., *Partial Observability in bivariate probit models*, Journal of econometrics, **12**, 1980, 209-217.