Astin Bulletin 12 (1981) 40-56

ÉTUDE STATISTIQUE DE LA PROBABILITÉ DE SINISTRE EN ASSURANCE AUTOMOBILE

MARC HALLIN et JEAN-FRANÇOIS INGENBLEEK
Institut de Statistique, Université Libre de Bruxelles.

Dans trois études récentes (1977a, 1977b, 1979), J. Lemaire a appliqué à un ensemble d'observations du risque automobile quelques méthodes de sélection fréquemment utilisées en analyse de la régression. Les variables *explicatives* (traitées comme variables *indépendantes*) sont les variables décrivant le risque. Les variables étudiées (traitées comme variables *dépendantes* d'un modèle de régression) sont les deux variables généralement prises en considération dans ce contexte: le nombre des sinistres et leur montant cumulé.

Nous avons déjà souligné (HALLIN, 1977) combien les hypothèses qui se trouvent à la base de l'analyse de la régression — normalité, homoscédasticité et linéarité de la régression — sont loin d'être remplies dans le contexte de l'assurance automobile. Le montant cumulé annuel prend la valeur zéro avec une probabilité proche de 0.9! Le nombre annuel de sinistres vaut 0, 1, 2, rarement plus! Même très approximativement, de telles variables peuvent difficilement être considérées comme normales. En outre, la plupart des variables explicatives sont de type nominal ou ordinal, ce qui rend délicate l'utilisation de modèles linéaires, les interactions de tous ordres étant très importantes, ainsi qu'on pourra le constater. Ces réserves sont d'ailleurs prévues par Jean Lemaire lui-même, qui ne propose ses conclusions qu'à titre de première approximation. Les mêmes données ont encore été soumises (MASURE, 1978) aux méthodes de l'analyse discriminante, et les mêmes réserves peuvent être faites en ce qui concerne l'utilisation des méthodes et l'interprétation des résultats (combinaisons linéaires, etc.).

Nous avons proposé dans HALLIN (1977a, b) un ensemble de méthodes qui, selon les hypothèses distributionnelles pouvant être faites (et qui vont des hypothèses classiques de l'analyse de la variance à celles, beaucoup moins restrictives, des méthodes de rangs), constituent des généralisations de celles qui sont utilisées par Jean Lemaire. En particulier, celle que nous appliquons ici est entièrement "*distribution free*". Ces méthodes sont également une extension de celle qu'a proposée PITKÄNEN (1975, 1976).

1. LES DONNÉES

Nous analysons donc ici une fois encore les données de Jean Lemaire, qui nous les a aimablement communiquées.

Un questionnaire a été rempli par 3879 souscripteurs sélectionnés au hasard

dans l'ensemble des souscripteurs d'une grande compagnie belge. Chacun de ces questionnaires porte les renseignements suivants:

- *nombre de sinistres en droit* [1]
- *âge du souscripteur* [2]
- *niveau de prime (dans l'échelle de bonus)* [2]
- *cylindrée du véhicule* [3]
- *âge du véhicule* [2]
- *prime effectivement payée* [2]
- *nombre de véhicules possédés par le souscripteur* [2]
- *nombre d'enfants du souscripteur* [2]
- *kilométrage annuel total moyen* [2]
- *kilométrage moyen parcouru pendant les vacances* [2]
- *kilométrage moyen parcouru pour le travail* [2]
- *distance habitation-travail* [2]
- *profession* [2]
- *nationalité* [2]
- *état civil* [2].

Un certain nombre de renseignements sont fournis par des variables dichotomiques:

- *usage tourisme-affaires / non*
- *usage mixte / non*
- *le souscripteur est sédentaire / non*
- *le souscripteur est de sexe masculin / non*
- *le souscripteur est de sexe féminin / non*
- *le souscripteur est une personne morale / non*
- *le souscripteur est de régime linguistique français / le souscripteur est de régime linguistique flamand*
- *le souscripteur habite une ville de plus de 5 000 habitants / non*
- *le souscripteur habite une ville de plus de 40 000 habitants / non*
- *le souscripteur seul est conducteur / les membres de sa famille conduisent également le véhicule assuré.*

Parmi les variables *explicatives* qui ne sont pas déjà dichotomiques, il convient de distinguer celles qui sont de type simplement *nominal* (état civil, nationalité, profession, ...) de celles qui ont un sens (au moins) *ordinal* (âge du conducteur, niveau de prime, cylindrée, ...). En vue de l'application de

[1] Au cours des 18 premiers mois de la période (trente mois) d'observation.
[2] Au début de la période d'observation
[3] Véhicule assuré au début de la période d'observation

notre procédure de sélection, il convient de traduire chacune de ces variables par un ensemble de variables dichotomiques. Le domaine de variation des variables ordinales a été découpé en un certain nombre de classes. Ainsi, pour la variable *âge du souscripteur*, sept classes ont été envisagées (les valeurs sont exprimées en nombres entiers d'années):

18-20 / 21-25 / 26-30 / 31-40 / 41-50 / 51-65 / 66 et plus

À ces sept classes correspondent six variables dichotomiques $X_1, \ldots, X_6$ définies par

$$X_i = \begin{cases} 0 \text{ si l'âge appartient à l'une des classes } i, i-1, \ldots, 1 \\ 1 \text{ si l'âge appartient à l'une des classes } i+1, i+2, \ldots, 7. \end{cases}$$

Un assuré d'âge 29 ans sera donc représenté par les valeurs

$$x_1 = x_2 = 1 \qquad x_3 = x_4 = x_5 = x_6 = 0.$$

Pour les variables purement nominales, le codage en variables dichotomiques est plus délicat. La variable état civil prend, par exemple, les modalités *célibataire-marié-veuf-séparé-divorcé*. Il existe quatorze partitions de cet ensemble de modalités en deux sous-ensembles propres; pour être complet, il faudrait par conséquent introduire quatorze variables dichotomiques. Pour des raisons de volume nous en avons sélectionné six:

$$X_1 = \begin{cases} 0 \text{ marié ou célibataire} \\ 1 \text{ sinon} \end{cases} \qquad X_2 = \begin{cases} 0 \text{ marié} \\ 1 \text{ sinon} \end{cases}$$

$$X_3 = \begin{cases} 0 \text{ marié, célibataire ou veuf} \\ 1 \text{ sinon} \end{cases} \qquad X_4 = \begin{cases} 0 \text{ marié ou veuf} \\ 1 \text{ sinon} \end{cases}$$

$$X_5 = \begin{cases} 0 \text{ séparé} \\ 1 \text{ sinon} \end{cases} \qquad X_6 = \begin{cases} 0 \text{ divorcé} \\ 1 \text{ sinon} \end{cases}$$

Par dichotomisation de toutes les variables explicatives, nous obtenons un ensemble de 89 variables dichotomiques; à chacune d'elles correspond une division en deux parties de l'échantillon. Dans la suite, nous les noterons $X_1, X_2, \ldots, X_i, \ldots, X_{89}$, réservant les minuscules $x_1, \ldots, x_{89}$ aux valeurs prises par ces variables.

2. LA MÉTHODE

Pour chacun des souscripteurs interrogés, on dispose également, bien entendu, du nombre et du montant cumulé des sinistres, et ce pour trente mois consécutifs. Ces deux variables admettent cependant des distributions qui se prêtent mal à une analyse statistique. Aussi nous semble-t-il préférable d'étudier séparément

- le montant cumulé des sinistres pour les assurés ayant un sinistre au moins
- la probabilité de sinistre (probabilité de causer un sinistre au moins).

C'est à cette probabilité que nous nous intéressons ici. Le nombre d'observations dont nous disposons est en effet trop faible pour qu'une étude du montant cumulé des sinistres puisse être entreprise de façon satisfaisante. Une modification de la statistique ϕ utilisée ci-dessous permettrait cependant cette étude, comme il est indiqué dans HALLIN (1977 b).

Supposons que k variables dichotomiques, notées $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$ soient prises en considération. Ces variables déterminent un découpage de l'échantillon en $m((1), (2), \ldots, (k))$ cellules non vides ($m \leqslant 2^k$) d'effectifs respectifs $n(x_{(1)}, \ldots, x_{(k)})$ ($x_{(i)} \in \{0, 1\}$). Soient d'autre part $p_0(x_{(1)}, \ldots, x_{(k)})$ et $n_0(x_{(1)}, \ldots, x_{(k)})$ les probabilités et les nombres de cas de non-sinistre dans ces cellules.

L'introduction d'une variable supplémentaire $X_{(k+1)}$ divise chacune des cellules existantes en deux sous-cellules, auxquelles correspondent les effectifs $n(x_{(1)}, \ldots, x_{(k)}, x_{(k+1)})$, $n_0(x_{(1)}, \ldots, x_{(k)}, x_{(k+1)})$, et les probabilités $p_0(x_{(1)}, \ldots, x_{(k)}, x_{(k+1)})$. Nous dirons qu'une cellule caractérisée par les valeurs $(x_{(1)}, \ldots, x_{(k)})$ des k variables de départ est *proprement divisée* par $X_{(k+1)}$ si $n(x_{(1)}, \ldots, x_{(k)}, 0)$ et $n(x_{(1)}, \ldots, x_{(k)}, 1)$ sont tous deux positifs; soit $l((k+1) \mid (1), (2), \ldots, (k))$ le nombre de cellules proprement divisées par $X_{(k+1)}$.

On peut considérer que chacun des effectifs $n_0(x_{(1)}, \ldots, x_{(k+1)})$ admet une distribution binomiale de paramètre $p_0(x_{(1)}, \ldots, x_{(k+1)})$ et d'exposant $n(x_{(1)}, \ldots, x_{(k+1)})$ (conditionnellement à n); sous l'hypothèse

$$H_0: p_0(x_{(1)}, \ldots, x_{(k)}, 0) = p_0(x_{(1)}, \ldots, x_{(k)}, 1) \quad \text{pour toute cellule } (x_{(1)}, \ldots, x_{(k)}) \text{ proprement divisée par } X_{(k+1)}, \tag{1}$$

la statistique

$$\phi((k+1) \mid (1), (2), \ldots, (k)) = \sum_{1}^{l} \left[\left(\frac{n_0(x_{(1)}, \ldots, x_{(k)}, 0)}{n(x_{(1)}, \ldots, x_{(k)}, 0)} - \frac{n_0(x_{(1)}, \ldots, x_{(k)}, 1)}{n(x_{(1)}, \ldots, x_{(k)}, 1)} \right)^2 \frac{n^3(x_{(1)}, \ldots, x_{(k)})}{n_0(x_{(1)}, \ldots, x_{(k)}) n_1(x_{(1)}, \ldots, x_{(k)})} \right],$$

https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press

où

$$n_1(x_{(1)}, \ldots) = n(x_{(1)}, \ldots) - n_0(x_{(1)}, \ldots)$$

(la somme $\sum_{i}^{I}$ s'effectuant sur les cellules proprement divisées par $X_{(k+1)}$) est asymptotiquement distribuée comme une variable χ^2 à $l((k+1) \mid (1), (2), \ldots, (k))$ degrés de liberté. De fait, nous avons préféré utiliser la transformation angulaire:

$$\phi((k+1) \mid (1), (2), \ldots, (k)) =$$

$$\sum_{i}^{I} \left[\left(2 \arcsin \sqrt{\frac{n_0(x_{(1)}, \ldots, x_{(k)}, 0)}{n(x_{(1)}, \ldots, x_{(k)}, 0)}} - 2 \arcsin \sqrt{\frac{n_0(x_{(1)}, \ldots, x_{(k)}, 1)}{n(x_{(1)}, \ldots, x_{(k)}, 1)}} \right)^2 \right.$$
$$\left. \Big/ \left(\frac{1}{n(x_{(1)}, \ldots, x_{(k)}, 0)} + \frac{1}{n(x_{(1)}, \ldots, x_{(k)}, 1)} \right) \right].$$

Ces statistiques permettent de tester l'hypothèse (1) contre l'hypothèse H_1 qu'il existe au moins une cellule proprement divisée donnant naissance à un couple de probabilités p_0 différentes.

La procédure de sélection (ou de segmentation) se déroule alors de la façon suivante. Les variables X sont sélectionnées une à une, par récurrence, selon le principe des méthodes du type "*pas à pas*" (*stepwise*) (cf. DRAPER and SMITH (1966)). Chaque étape de la méthode comporte deux parties distinctes: introduction de la variable dont la contribution semble la plus significative (conduisant le plus nettement au rejet de (1)), puis élimination éventuelle d'une variable devenue non significative.

Étape k.

(*k1. phase d'introduction*):

Notons $X_{(1)}, \ldots, X_{(k-1)}$ les variables obtenues à la fin de l'étape précédente. Remarquons que cet ensemble peut comporter un nombre de variables strictement inférieur à $k-1$ et que, en dépit de la notation, $X_{(1)}$, première variable sélectionnée, peut n'en plus faire partie. Pour chacune des variables X_i restantes, considérons les valeurs ϕ_i prises par $\phi(i \mid (1), \ldots, (k-1))$; à chacune des ces quantités correspond un niveau de signification q_i, valeur en ϕ_i de la fonction de répartition d'une variable χ^2 à $l(i \mid (1), \ldots, (k-1))$ degrés de liberté. Soit $q_{(k)}$ le plus élevé de ces niveaux de signification: $X_{(k)}$ est, provisoirement, la $k^{\text{ième}}$ variable sélectionnée.

(*k2. phase d'élimination*):

Considérons à présent, pour chacune des variables $X_{(l)}$ sélectionnées ($X_{(k)}$ comprise), la valeur $\phi_{(l)}$ prise par $\phi((l) \mid (1), \ldots, (l-1), (l+1), \ldots, (k))$. A chacune de ces valeurs correspond à nouveau un niveau de signification $q_{(l)}$. Soit q_m le plus bas de ces niveaux:

(*k2a*): si $q_m > 1 - \alpha$, on passe à l'étape $k+1$ avec $\{X_{(1)}, \ldots, X_{(k)}\}$ pour nouvel ensemble de variables sélectionnées (α étant un niveau de probabilité fixé à l'avance).

(*k2b*): si $q_m \leqslant 1 - \alpha$, l'hypothèse

H_0: $p_0(x_{(1)}, \ldots, x_{(m-1)}, 0, x_{(m+1)}, \ldots, x_{(k)}) = p_0(x_{(1)}, \ldots, x_{(m-1)}, 1, x_{(m+1)}, \ldots, x_{(k)})$ $\forall$ $x_{(1)}, \ldots, x_{(m-1)}, x_{(m+1)}, \ldots, x_{(k)}$

ne peut être rejetée au niveau α; si $m \neq k$, on passe à l'étape $k+1$ avec $\{X_{(1)}, \ldots, X_{(m-1)}, X_{(m+1)}, \ldots, X_{(k)}\}$ pour nouvel ensemble de variables; si $m = (k)$, la procédure s'arrête, l'ensemble final étant $\{X_{(1)}, \ldots, X_{(k-1)}\}$.

Le cas de la variable *personne physique/personne morale* doit être considéré séparément, une "personne morale" n'ayant ni sexe, ni nombre d'enfants, ni état civil, etc. Aussi cette distinction doit-elle être introduite automatiquement dès que l'une des variables "personnalisées" (*sexe, état civil, nombre d'enfants, kilomètres vacances*, ...) est sélectionnée, et indépendamment de son niveau de signification. En outre, lors du calcul, en cours d'étape, de la valeur prise par la statistique ϕ relative à l'une de ces variables "personnalisées", les "personnes morales" doivent être soigneusement omises.

Cette méthode a été programmée par J.-F. Ingenbleek pour une CDC 6600. Les procédures prévues pour les cas de valeurs manquantes et les cellules trop peu peuplées ont été améliorées par rapport à une version précédente du programme (Hallin et Ingenbleek, 1979). Pour obtenir les résultats qui suivent, nous n'avons considérées comme *proprement divisées par une variable* X_k que les cellules donnant naissance, du fait de la valeur 0 ou 1 prise par X_k, à deux cellules d'effectif supérieur ou égal à 15 (l'effectif de la cellule d'origine étant donc supérieur à 30); les cellules trop peu peuplées n'entrent ainsi pas en ligne de compte dans le calcul des statistiques $\phi(. \mid \ldots)$. Nous avons appliqué aux valeurs manquantes le traitement suivant. Supposons que la valeur d'une variable X_i soit inconnue pour un assuré, celui-ci ayant mal rempli le questionnaire qui lui a été soumis. Si X_i ne figure pas dans l'ensemble des variables sélectionnées en début d'étape, cet assuré constitue, pour le calcul, lors de la phase d'introduction, de tous les $\phi(j \mid \ldots)$, $j \neq i$, une observation parfaitement valide. En revanche, lorsque le programme en vient à envisager l'introduction éventuelle, dans le tarif, de X_i et calcule donc $\phi(i \mid \ldots)$, cet assuré est omis, et il doit être tenu compte de cette omission dans l'obtention du niveau de probabilité correspondant. Au cas où X_i figurerait dans l'ensemble des variables déjà sélectionnées en début d'étape, le même assuré, ne pouvant être classé en fonction des variables en tarif, doit être omis dans *tous* les calculs, et ce jusqu'à l'élimination éventuelle de X_i.

Ces modifications et le nombre, hélas élevé [4], des valeurs manquantes pour

[4] Pour l'*âge du souscripteur* et l'*âge du véhicule assuré*, ce nombre atteint près du tiers de la taille de l'échantillon!

certaines variables explicatives expliquent les différences de résultats entre les deux versions.

3. COMMENTAIRES

3.1. L'application de techniques du type "analyse de la variance" à des tables de contingence (variables dépendantes de type binomial ou multinomial) soulève toujours un grand nombre de problèmes, surtout lorsque les fréquences varient, comme c'est le cas ici, de cellule à cellule. Même le cas le plus simple et le plus classique de la comparaison de deux proportions ne peut être traité (cf. GART, 1971) de façon uniformément satisfaisante.

De nombreuses variantes aux méthodes classiques, reposant sur des choix de pondérations et de transformations de variables, ont été proposées (COCHRAN, 1943 et 1954, GART, 1971, ...). En l'absence de *modèle* liant les proportions observées aux variables explicatives, il est cependant impossible d'opérer un choix parmi ces méthodes, ni même de faire appel à la notion de puissance locale. Or, dans le cas qui nous occupe, les variables sont beaucoup trop nombreuses, les multicolinéarités et les interactions beaucoup trop considérables, pour qu'un modèle à la fois simple et réaliste puisse être construit. Le choix de la statistique ϕ sur laquelle repose la sélection est donc en grande partie arbitraire, la notion même de "meilleure statistique" n'ayant pas de sens.

Nous avons négligé, en outre, les phénomènes de *variations étrangères* (*extraneous variations* — cf. COCHRAN, 1943), nous bornant à considérer les observations comme engendrées par des processus binomiaux purs. Le niveau de probabilité des tests effectués peut présenter par conséquent certaines distorsions; il est plus prudent de se fixer une valeur de α assez faible ($\alpha = 1\%$, par exemple).

3.2. Indépendamment du choix de la statistique ϕ utilisée, notre méthode souffre d'un certain nombre de défauts inhérents à toutes les procédures de type *stepwise* communément utilisées. Les tests non indépendants, effectués en chaîne, conduisent à un niveau global difficile à apprécier; pour certaines étapes, le nombre élevé de cellules dépeuplées provoque, par perte de degrés de liberté, un amenuisement de la quantité d'information contenue dans la statistique χ^2. Pire: en présence d'un large éventail de variables explicatives, et en raison des multicolinéarités et des interactions inévitables, ces procédures débouchent, le plus souvent, sur un cyclage (c'est notamment le cas ici pour $\alpha = 1\%$).

Ces réserves d'ordre théorique ont cependant peu de répercussions sur les applications, et ne doivent pas masquer la richesse des renseignements fournis à chaque étape [5]. De toute façon, comme nous l'avons déjà souligné (HALLIN,

[5] On lira avec intérêt, à ce sujet, les commentaires qui accompagnent l'étude de la byssinose respiratoire chez les travailleurs de l'industrie cotonnière (HIGGINS et KOCH, 1977).

1977), le problème posé (celui de la recherche du "meilleur" sous-ensemble de variables explicatives) est un problème mal posé; aucun critère permettant de classer entre eux les divers sous-ensembles possibles ne s'impose de façon absolue. Et, quand bien même un tel critère existerait, la variation, selon l'échantillon considéré, du sous-ensemble sélectionné, est un phénomène essentiellement non quantifiable. Toute procédure de sélection, que ce soit dans le cadre d'une analyse de la régression ou dans le cadre plus général que nous considérons ici, doit être appliquée de façon assez heuristique, comme une méthode "applicable", fournissant des ensembles "intéressants" de variables explicatives. Et les résultats intermédiaires aussi bien que les résultats finals doivent être examinés dans une optique d'*analyse de données*.

4. LES RÉSULTATS

4.1. *Au niveau de probabilité de 0,5%*

Au niveau de probabilité $\alpha = 0{,}5\%$, la procédure s'arrête après dix étapes.

TABLEAU 1

Étape	Variable entrante	Variable sortante
1	*niveau de prime* [1]: moins de 80%/80% et plus	—
2	*niveau de prime* [1]: moins de 70%/70% et plus	—
3	*zone de garage*: moins de 40 000 hab./plus	—
4	*personne morale/non*	—
5	*cylindrée*: moins de 900 cc/plus	*niveau de prime* [1]: moins de 70%/70% et plus
6	*niveau de prime* [1]: moins de 65%/65% et plus	*zone de garage*: moins de 40 000 hab./plus
7	*kilométrage annuel*: moins de 10 000 km/an/plus	—
8	*profession*: commerçant, ouvrier, employé, cadre/autres	—
9	*niveau de prime* [1]: moins de 70%/70% et plus	—
10	*zone de garage*: moins de 40 000 hab./plus	*zone de garage*: moins de 40 000 hab./plus STOP

[1] Exprimé en pourcentage de la prime totale.

https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press

Le Tableau 2 ci-dessous donne les 7 variables explicatives finalement sélectionnées; pour chacune de ces variables, on indique

— la valeur de la statistique ϕ permettant de tester la "sortie" éventuelle de cette variable
— le nombre de degrés de liberté de la distribution de cette statistique
— le niveau de signification (probabilité laissée "à gauche" sous l'hypothèse nulle).

Afin de ne pas accorder une influence excessive aux cellules de faible fréquence, un effectif minimum de quinze observations a été exigé pour qu'une cellule soit prise en considération dans le calcul de ϕ. Si donc une variable découpe en deux sous-cellules d'effectif supérieur ou égal à quinze l'une des cellules construites sur les autres variables, cette division apporte un degré de liberté à la statistique ϕ. Ainsi, les personnes morales n'étant pas très nombreuses dans l'échantillon, la statistique correspondant à cette variable ne jouit-elle que d'un seul degré de liberté.

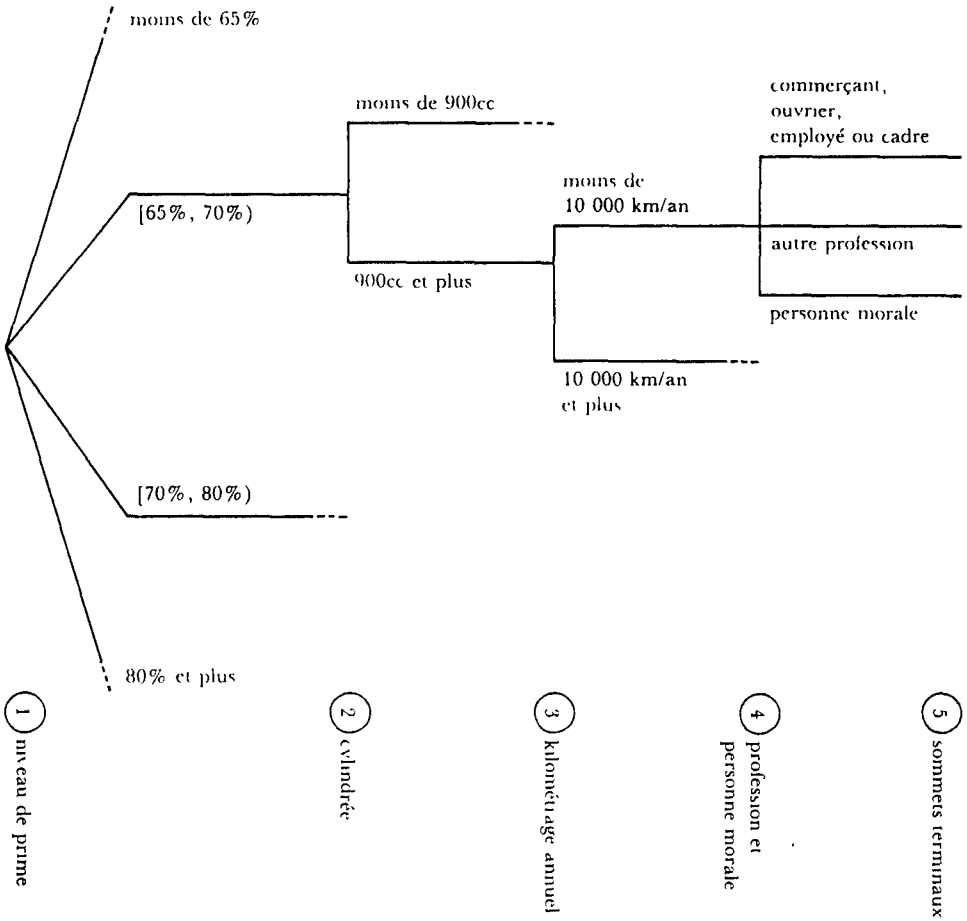
TABLEAU 2

Variable	Statistique ϕ	Degrés de liberté	Niveau de signification
niveau de prime: 65%	31,9802	5	1,000
70%	19,4802	5	0,9984
80%	44,0269	8	1,0000
cylindrée: 900 cc	36,1261	13	0,9994
kilométrage annuel: 10 000 km/an	37,1952	14	0,9993
profession: commerçant, ouvrier, employé, cadre/autres	37,0007	14	0,9993
personne morale/non	0,1589	1	0,3098

Ces sept variables découpent théoriquement dans l'ensemble des assurés 48 cellules distinctes. Certaines de ces cellules (niveau de prime compris entre 65% et 70% et cylindrée inférieure à 900 cc) étant peu peuplées, nous en avons retenu 41. Le graphe ci-dessous donne, pour chacune de ces 41 cellules (représentées par les sommets terminaux de l'arborescence), le nombre n d'observations, le nombre n_1 de cas présentant un sinistre au moins, et, lorsque n est suffisamment élevé, l'estimation $\hat{p} = n_1/n$ de la probabilité de sinistre (d'un sinistre au moins *sur trente mois consécutifs*).

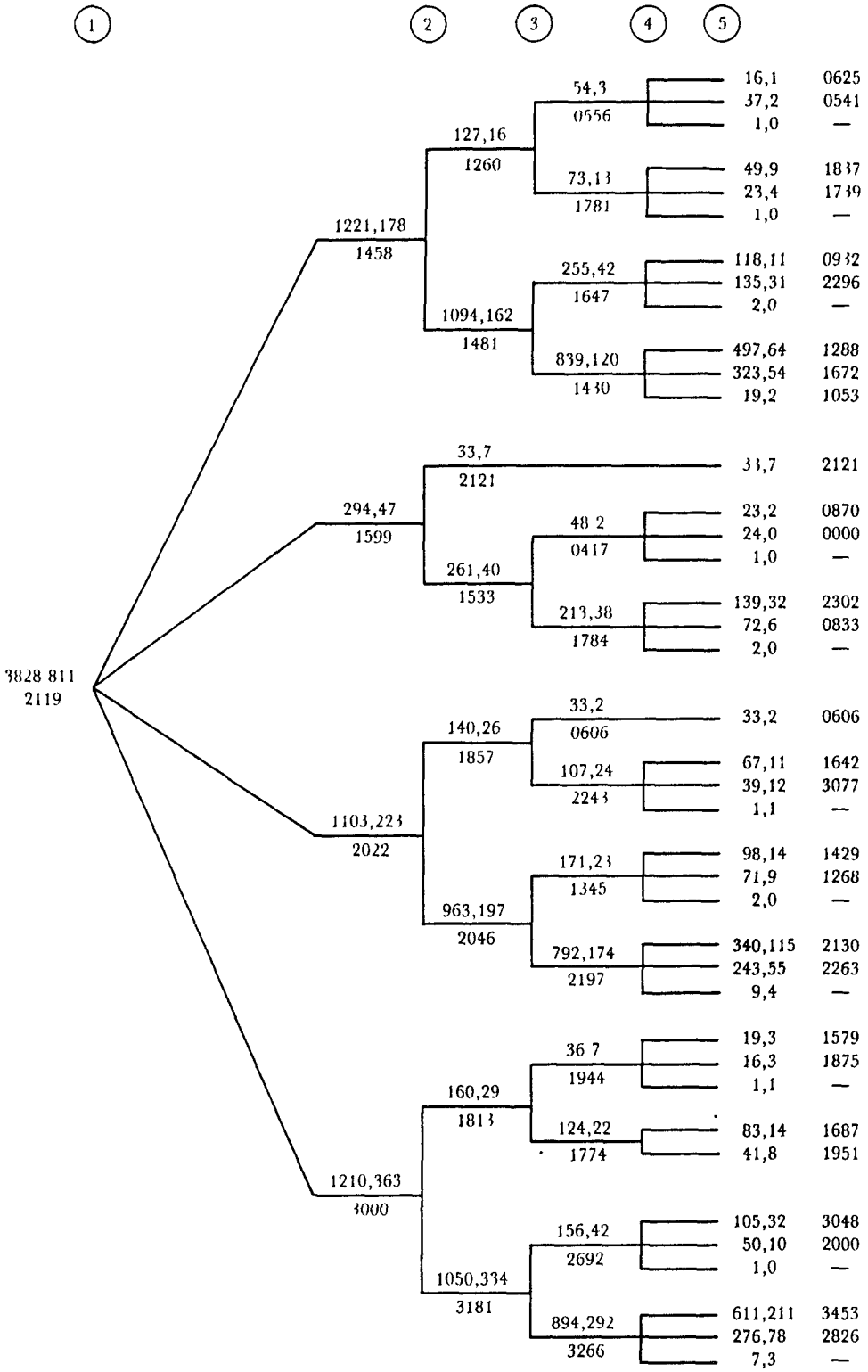
Le schéma suivant indique comment doit être interprétée l'arborescence (pour la construire il a bien fallu attribuer un ordre plus ou moins arbitraire aux variables sélectionnées).

Les nombres qui accompagnent chacun des sommets intermédiaires sont n, n_1, et $\hat{p} = n_1/n$.



Commentaires

Toutes les variables sélectionnées sont *très* significatives. Le *niveau de prime*, en particulier, présente de très bonnes performances, puisqu'il détermine 4 classes d'assurés. Il semblerait cependant que le "bas" de l'échelle (de 60% à 80%) gagnerait à être raffiné, tandis que, dans le "haut" de l'échelle, une distinction entre les conducteurs de niveau 80% et les conducteurs de niveau 120%, par exemple, ne paraît pas très justifiée. Il en est de même pour les autres variables retenues: *cylindrée* et *kilométrage annuel*. Des distinctions très fines ne semblent pas s'imposer, et une séparation entre les petites cylindrées et les moyennes et grosses (900 cc et plus), entre les faibles kilométrages et les moyens et gros kilométrages (10 000 km et plus) apparaît comme largement suffisante.



https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press

L'ordre d'entrée et la sortie éventuelle des variables indique également les dépendances et les interactions: l'introduction (étape 5) de la *cylindrée* provoque le remplacement du *niveau de prime* 70% par le *niveau* 65% (étape 6); et ce dernier "chasse" la *zone de garage* au profit du *kilométrage annuel* (étape 7): si les kilomètres parcourus en ville sont plus fertiles en accrochages, le niveau de prime en tient suffisamment compte pour que la distinction entre kilomètres urbains et non urbains soit superflue. On remarque également que l'effet néfaste de la cylindrée et des kilomètres s'exerce de façon beaucoup plus importante chez les "mauvais" conducteurs (80% et plus) que chez les "bons".

Les meilleurs risques sont observés, comme on peut s'y attendre, dans le haut du graphe: "bons" conducteurs, roulant peu dans une voiture de petite cylindrée: $\hat{p} = 0{,}0556$. Les plus mauvais risques, au bas du graphe, avec $\hat{p} = 0{,}3266$ ($n_1/n = 292/894$, ce qui donne un intervalle de confiance assez bon, au niveau de 5%: [0,2959; 0,3573]).

On pourrait ainsi multiplier les commentaires: il suffit d'examiner le graphe. Il convient cependant de rester prudent: l'estimateur $\hat{p}$ n'a pas une variance négligeable, même pour un nombre relativement élevé d'observations.

4.2. *Au niveau de probabilité de 1%*

Au niveau de probabilité de 1%, la variable *zone de garage* ne ressort plus à la 10e étape, et la procédure se poursuit de la façon suivante (Tableau 3).

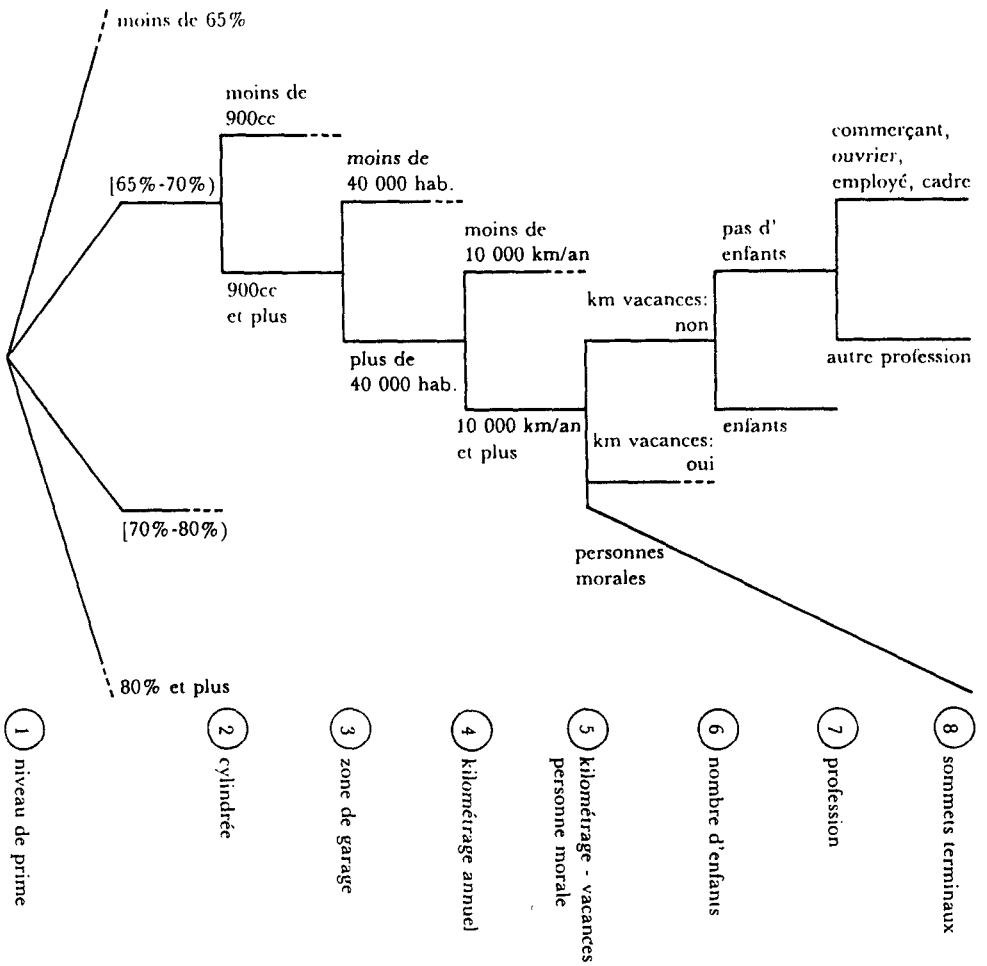
TABLEAU 3

Étape	Variable entrante	Variable sortante
10	*zone de garage*: moins de 40 000 hab./plus	—
11	*nombre d'enfants*: 0, 1, 2/3 et plus	—
12	*nombre d'enfants*: 0/1 au moins	*niveau de prime*: moins de 70%/70% et plus
13	*kilométrage vacances*: 0/1 km au moins	*nombre d'enfants*: 0, 1, 2/3 et plus
14	*zone de garage*: moins de 5 000 hab./plus	*kilométrage annuel*: moins de 10 000 km/an/plus
15	*niveau de prime*: moins de 70%/70% et plus	*kilométrage vacances*: 0/1 km au moins
16	*nombre d'enfants*: 0, 1, 2/3 et plus	*zone de garage*: moins de 5 000 hab./plus
17	*kilométrage annuel*: moins de 10 000 km/an/plus	*niveau de prime*: moins de 70%/70% et plus
.		
.		
.		

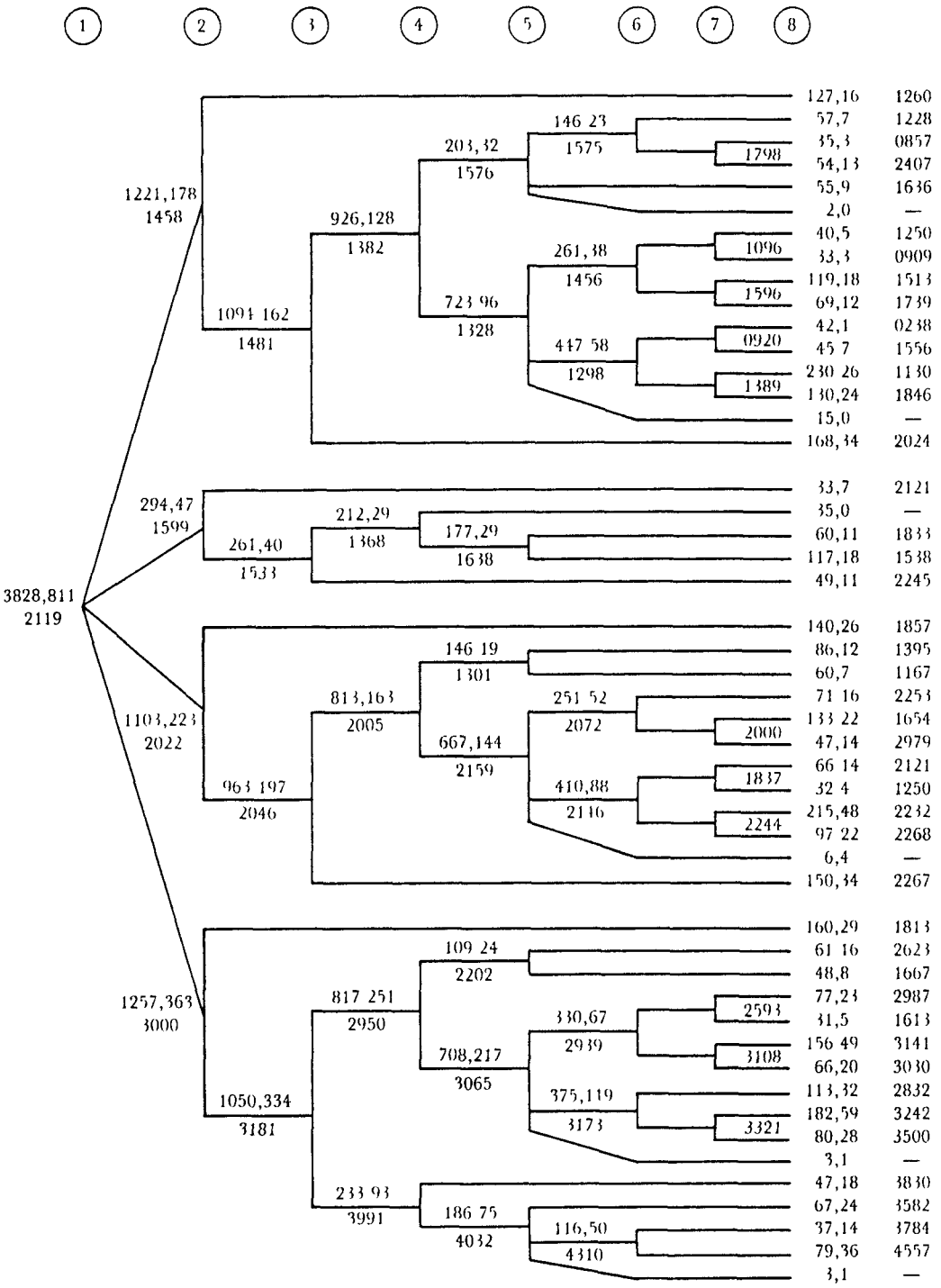
https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press

A la sortie de l'étape 17, la situation est la même qu'au début de l'étape 13, ce qui entraîne la procédure dans un cycle de période 5 étapes; les sélections de variables correspondant à ces cinq étapes présentent des qualités assez semblables.

A titre d'illustration, nous avons choisi de présenter, pour l'une des étapes du cycle, une arborescence équivalente à celle que nous avons donnée pour l'étape 10. Le schéma ci-dessous indique la façon de lire cette arborescence.



https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press



https://doi.org/10.1017/S0515036100006814 Published online by Cambridge University Press

4.3. *Autres résultats.* Comparaison avec les conclusions de JEAN LEMAIRE (1979).

Ici encore, il est intéressant d'observer la façon dont les variables s'introduisent et se "chassent" mutuellement. Nous n'avons pas effectué l'analyse factorielle ou en composantes principales appropriée de ces données. On peut cependant, à l'examen des étapes 12 à 17, se risquer à discerner, derrière les diverses variables qui interviennent, trois types d'effets ou de *facteurs*: l'un — disons f_1 — mesure l'intensité d'exposition au risque du véhicule assuré (et n'est pas forcément proportionnel au kilométrage annuel moyen); un second — disons f_2 — est lié à l'environnement (plus ou moins urbain) dans lequel est utilisé le véhicule; le troisième enfin caractérise l'attitude au volant du conducteur du véhicule. Chacune des variables apparaissant au cours des étapes 12 à 17 peut être considérée comme un index plus ou moins représentatif de ces trois effets: le nombre d'enfants est essentiellement lié à f_3 (conduite de "père de famille"), mais aussi à f_1; la zone de garage à f_2 et f_1, etc. Ceci explique que le *nombre d'enfants* "chasse" la *zone de garage* au profit du *kilométrage annuel*, et que, à la sortie du *nombre d'enfants*, la *zone de garage* revienne se substituer au *kilométrage annuel* ...

Outre la sélection des variables, notre programme fournit un grand nombre de renseignements concernant les variables non sélectionnées. Pour chaque ensemble de variables $X_{(1)}, X_{(2)}, \ldots, X_{(k)}$ considéré en début d'étape, et pour chaque variable $X_i \neq X_{(1)}, \ldots, X_{(k)}$ on dispose des effectifs $n(x_{(1)} \ldots x_{(k)} x_i)$ et $n_1(x_{(1)} \ldots x_{(k)} x_i)$, des estimations $\hat{p}(\ldots) = n_1(\ldots)/n(\ldots)$, de la statistique $\phi\ (i \mid (1) \ldots (k))$, de son nombre de degrés de liberté et de son niveau de signification (probabilité à gauche sous H_0).

Ainsi, lors de la première étape (ensemble sélectionné en début d'étape: Φ), les variables suivantes sont significatives à 1%:

— nombre de sinistres en droit (0/1 ou plus)
— niveau de prime (quatre valeurs: 65%, 70%, 80% et 90%; c'est 80% qui sera sélectionné)
— zone de garage
— âge du souscripteur (quatre valeurs: 26, 31, 41 et 51 ans!)
— kilométrage annuel (5 000, 10 000 et 15 000 km/an)
— distance habitation-travail (10 km)
— état civil (mariés/autres; mariés et veufs/autres).

On remarquera l'absence, à ce niveau, et très significativement, des variables *usage tourisme et affaires* (niveau de signification 0,23), *souscripteur sédentaire* (dans le tarif actuel, donne droit à une réduction de prime de 15%, niveau de signification 0,54!), *sexe* (niveau 0,77), *nationalité*, *profession*, *âge du véhicule*.

La *cylindrée* n'est présente qu'avec une seule valeur, 900 cc, qui sera sélectionnée à l'étape 3; pour 1100 cc, le niveau de signification tombe à 0,21.

Ceci semble bien indiquer que la "taille" de la voiture agit à la façon d'une variable dichotomique (petites voitures/autres), non à la façon d'un régresseur linéaire (la nature de cette régression, d'une variable de type binomial en une variable continue, n'étant guère précisée, d'ailleurs, chez Jean Lemaire). En outre, l'introduction du *niveau de prime* met en évidence une interaction:

	cylindrée < 900 cc	cylindrée ⩾ 900 cc
niveau de prime < 80%	$\hat{p} = 0{,}1617$	$\hat{p} = 0{,}1720$
niveau de prime ⩾ 80%	$\hat{p} = 0{,}1786$	$\hat{p} = 0{,}3177$

La *cylindrée* n'a donc pas d'effet notable pour les "bons" conducteurs.

La troisième étape fournit les mêmes renseignements, mais en tenant compte de trois classes de niveau de prime (moins de 70%; [70%-80%); 80% et plus). Un grand nombre des variables qui étaient significatives lors de la première étape ne le sont plus: *nombre d'accidents en droit*, autres niveaux de prime (tous au-dessous d'un niveau de signification de 0,68, ce qui indique bien que l'échelle de bonus utilisée est probablement d'une complexité inutile), *distance habitation-travail*, *état civil*, *âge du souscripteur*.

Il est intéressant, à cet égard, de remarquer que, si, à l'étape 1, la valeur la plus significative de la variable *âge du souscripteur* est 26 ans ($\hat{p} = 0{,}3292$ pour les moins de 26 ans, $\hat{p} = 0{,}2253$ pour les plus de 26 ans), l'introduction d'un seul niveau de prime (80%) suffit à déplacer cette valeur à 41 ans (niveau de signification: 0,99):

	âge < 41 ans	âge ⩾ 41 ans
niveau de prime < 80%	$\hat{p} = 0{,}1967$	$\hat{p} = 0{,}1934$
niveau de prime ⩾ 80%	$\hat{p} = 0{,}3389$	$\hat{p} = 0{,}2375$

Si, par conséquent, les "jeunes" constituent un moins bon risque que les "moins jeunes", l'utilisation d'une échelle de bonus-malus, même rudimentaire, suffit à en rendre compte. La franchise de 4000 FB qui, dans le tarif actuel, est systématiquement infligée à tout conducteur de moins de 23 ans ne se justifie donc absolument pas. Il est également intéressant de noter que, pas plus que la cylindrée ni le kilométrage, l'âge n'a d'effet important sur les "bons" souscripteurs.

En conclusion, le *niveau de prime*, surtout du côté de ses basses valeurs, confirme ses qualités d'excellent critère de discrimination entre "bons" et "moins bons" risques. Seuls conservent intacte leur significativité la *cylindrée* (toujours à 900 cc), la *zone de garage* et le *kilométrage annuel*.

Il faut souligner, toutefois, que ces remarques et ces conclusions sont relatives à la probabilité de sinistre uniquement. Il est tout à fait possible, et

même probable, qu'un examen des montants cumulés mène à des résultats fort différents: coût moyen des sinistres plus élevé chez les jeunes, dans les campagnes, chez les conducteurs faisant peu de kilomètres et transportant dans leur véhicule une nombreuse famille, etc. Malheureusement, comme nous l'avons dit plus haut, le nombre de sinistres observés dans l'échantillon dont nous disposons est trop peu élevé pour qu'une étude sérieuse puisse en être faite.

Nos résultats et ceux de Jean Lemaire, dans la mesure où ils peuvent être comparés, divergent essentiellement sur trois points: *l'âge du souscripteur*, sa *nationalité* et son *état civil*, sélectionnés chez Jean Lemaire, font place au *nombre d'enfants* (du moins à *pas d'enfant/un enfant au moins* — fortement lié à la variable *état civil*) et *kilométrage vacances* (encore une variable à caractère nettement dichotomique: *pas de km-vacances/1 km au moins*). Mais il ne faut pas oublier que, pour chacune de ces variables, chez Jean Lemaire, *tous* les niveaux sont testés et sélectionnés globalement.

A aucun moment, la *prime effectivement payée*, donc le tarif actuellement en vigueur, n'approche le seuil de significativité.

REFERENCES

COCHRAN, W. G. (1943). Analysis of variance for percentages based on inequal numbers. *JASA*, **38**, 287-301.

COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**, 417-451.

DRAPER, N. and H. SMITH (1966). *Applied Regression Analysis*. Wiley, N.Y.

GART, J. J. (1971). The comparison of proportions: a review of significance tests, confidence intervals and adjustements for stratification. *Review of the International Stat. Institute*, **39**, 148-69.

HALLIN, M. (1977a). Méthodes Statistiques de Construction de Tarif, *Bulletin de l'Association des Actuaires Suisses*, 162-175.

HALLIN, M. (1977b). Étude statistique des facteurs influençant un risque, *Bulletin de l'Association R. des Actuaires Belges*, 76-92.

HALLIN, M. et J.-F. INGENBLEEK (1978). Étude statistique des Facteurs influençant le Risque automobile: la probabilité de sinistre. *Discussion paper n° 5*, Institut de Statistique de l'Université Libre de Bruxelles.

HIGGINS, J. E. and G. G. KOCH (1977). Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review*, **45**, 51-62.

LEMAIRE, J. (1977a). Selection Procedures of Regression Analysis applied to Automobile Insurance, *Bulletin de l'Association des Actuaires Suisses*, 143-160.

LEMAIRE, J. (1977b). Critique du tarif automobile responsabilité civile belge. *Bulletin de l'Association R. des Actuaires Belges*, 93-109.

LEMAIRE, J. (1979). Selection Procedures of Regression Analysis applied to Automobile Insurance, Part II: Sample inquiry and underwriting applications. *Bulletin de l'Association des Actuaires Suisses*, 65-71.

MASURE, L. (1978). L'analyse discriminante appliquée aux problèmes de l'assurance automobile. *Bulletin de l'Association R. des Actuaires Belges*, 29-51.

PITKÄNEN, P. (1975). Tariff theory, *Astin Bulletin*, 204-228.

PITKÄNEN, P. (1976). *A theoretical approach to premium rating*. Int. Congress of Actuaries, Tokyo, 247-252.