

SOMMAIRE

SOMMAIRE.....	1
LISTE DES TABLEAUX ET GRAPHIQUES.....	3
DEDICACES.....	4
REMERCIEMENTS.....	5
AVANT-PROPOS.....	6
RESUME.....	7
ABSTRACT.....	7
PRESENTATION DE CHANAS ASSURANCES SA.....	8
INTRODUCTION.....	9

GENERALITES SUR L'ANALYSE DES COMPOSANTES MULTIPLES ET LA REGRESSION LOGISTIQUE

I – Analyse des correspondances multiples.....	11
A- Analyse générale.....	11
B- Analyse des correspondances.....	13
C- Analyse des correspondances multiples.....	15
II – Régression logistique.....	18
A –Introduction.....	18
B – Le modèle Logit.....	19
C – Estimations et tests.....	19

DESCRIPTION DES VARIABLES ET CODAGE

I – Principe de fonctionnement de l'assurance automobile.....	21
II – Méthodologie de collecte de données.....	22
A – Définition de « haut sinistré ».....	22
B – Description et codage des variables.....	23

DESCRIPTION MULTIDIMENSIONNELLE DES DONNEES

I – Analyse des correspondances multiples.....	26
A – Conditions d'utilisation de la méthode.....	26
B – Recodage de variables.....	28
C – Principe de l'analyse.....	29
D – Interprétations.....	30
II – Conclusion.....	37

MODELISATION

I – Pourquoi le modèle de régression logistique ?.....	38
A – Exigence du modèle.....	38
B – Méthodologie de l'analyse.....	39
C – Estimation du modèle.....	40
D – Validation du modèle.....	43
- Evaluation du pouvoir prédictif du modèle.....	44
- Choix de la probabilité seuil (S_0).....	45
- Erreur de prédiction.....	47
- Règle de décision finale.....	48
- Programme R de classification automatique.....	48
II – Conclusion.....	49
CONCLUSION GENERALE.....	50
PERSPECTIVES ET RECOMMANDATIONS.....	51
ANNEXES.....	52
- Principaux programmes R utilisés	
- Listings des résultats D'ACM (Logiciel SPAD 4.01)	
BIBLIOGRAPHIE.....	55

LISTE DES TABLEAUX ET GRAPHIQUES

Tableaux

Tableau 0-1 : Eléments de base de l'analyse des correspondances.....	14
Tableau 0-2 : Eléments de construction de l'analyse des correspondances.....	14
Tableau 2-1 : Statistiques élémentaires des données brutes.....	27
Tableau 2-2 : Histogramme des valeurs propres.....	31
Tableau 2-3 : Coordonnées, Contributions et Cosinus carrés.....	32
Tableau 2-4: Récapitulatif des modalités bien représentées et à bonne contribution.....	32
Tableau 2-5 : Valeurs-tests des modalités significatives de variables illustratives.....	33
Tableau 3-1 : Analyse bivariées.....	39
Tableau 3-2 : Analyse de colinéarité.....	41
Tableau 3-3 : Fréquence des probabilités estimées par le modèle sous (H_0).....	46
Tableau 3-4: Fréquence des probabilités estimées par le modèle sous (H_1).....	47

Graphiques

Graphique 2-1 : Carte des modalités (axes 2 et 3).....	34
Graphique 2-2 : Carte des modalités (axes 1 et 3).....	34
Graphique 2-3 : Carte des modalités (axes 3 et 4).....	35
Graphique 2-4 : Carte des modalités (axes 3 et 6).....	35
Graphique 2-5 : Carte des modalités (axes 3 et 5).....	36
Graphique 3-1 : Graphes de diagnostic du modèle 1.....	42
Graphique 3-2 : Courbe ROC.....	45
Graphique 3-3 : Histogramme des probabilités estimées.....	45

DEDICACES

Qu'il me soit permis de dédier ce modeste travail à :
Mes Parents :

- M. TCHATCHUENG FOKOUO Emile
- Mme TCHATCHUENG née KOM Pauline
 - Mme NEM née NOUMSI Anne

Mes Frères et Sœurs

REMERCIEMENTS

La réalisation de ce travail a été possible grâce au concours de nombreuses personnes auxquelles nous témoignons ici notre gratitude.

Nous pensons ici :

- ✓ A Monsieur **Alain VEILLE**, Directeur de CHANAS ASSURANCES S.A, agence de Yaoundé qui a bien voulu nous donner l'opportunité de mettre à profit certaines théories étudiées durant notre formation.
- ✓ A Monsieur **Jean-Grattien ZANOVI**, Directeur Général de l'Institut International des Assurances

- ✓ **A tout le personnel de la Compagnie CHANAS**, en particulier :

Messieurs : BELINGA Alain, MBIDA Léon, NSOE Faustin, NOUMEMEN Isaac, BANGWEN Jean-Yves, EYILI Hervé

Pour l'accueil chaleureux, leur soutien inébranlable et leur bonne volonté à m'enseigner le principe de fonctionnement du portefeuille automobile.

- ✓ A tout ceux qui de près ou de loin ont contribué au bon déroulement du Master de statistique ; nous pensons ainsi aux enseignants qui n'ont ménagé aucun effort pour la réussite de cette formation par leurs enseignements de qualité exceptionnelle. Il s'agit principalement du :

Pr. Didier DACUNHA CASTELLE
 Pr. Jean COURSOL
 Pr. Jean CHRISTOPHE THALABARD
 Pr. Jean LOUIS GOLMARD
 Pr. Bertran AUVERT
 Pr. Xavier GUYON
 Pr. Elisabeth GASSIAT
 Pr. Danielle FLORENS
 Pr. Henri GWET
 Dr. Eugène NDONG NGUEMA
 Dr. Michel NDOUMBE NKENG
 Dr. Maxime KIKI

- ✓ **A toute ma famille**, qui d'une manière ou d'une autre a apporté sa contribution à la réalisation de ce travail. Nous pensons ici aux familles :

KAMARA Abdoulaye, OMOKOLO Denis, NEM Joseph, KAMTCHUENG Célestin, TAFFE Polycarpe, MOGO Amos, KOM Pierre, KAMSU Jean-Paul, TAGNE Jean-Paul, KAMSU Duclos, KOMGUEM Depeskido, NZIETCHUENG Samuel, TACHOM Ernest, BOGNE Patrice, FOWA, NOUMSI, DIPEUH Alain, DJOUM Alain, NZIETCHUENG Bertin, SIGNE Pierre, KAMKUI Patrice et aux grandes familles Bù TATCHUENLIEU, KAMKUI Engelbert, MATAGNE et WABO TAMETCHA

- ✓ **A mes frères et amis** : Francis SIKADI, Cédric NOUMSI, Guillaume NZUKAM, Olivier MBIELEU, Justin SIEPEING, Christian FOTSING, Godwin FOMEN, Martial NEMPE, Donatien WAKAM, Engeline TCHOUOBIAP, Irénée DOMKAM, Rolin SILA, Herbert NGOWA, Rostand DOUANLA, Cyrille CHENKEM, Hervé-lys KWADJO, Hervé MOMEYA.

AVANT PROPOS

Le **Master de Statistique Appliquée** est une formation de troisième cycle ouverte et animée à l'université de Yaoundé I. Il s'agit d'une formation professionnalisante et d'initiation à la recherche bénéficiant de la collaboration et du soutien des Universités : Paris Orsay, Paris-Dauphine, Paris5, Versailles, Institut National Polytechnique-HB(côte d'Ivoire), INSERM (France).

L'objectif général de cette formation est de donner aux étudiants, cadres supérieurs d'entreprises et d'administrations, et tout utilisateur de la statistique, une formation de haut niveau très concrète, classique quant aux techniques mathématiques utilisées, aussi moderne que possible quant à l'informatique et aux logiciels spécialisés utilisés. Ce Master apporte aux étudiants ayant les acquis fondamentaux en Mathématiques et en Statistiques, une formation professionnelle complémentaire dans le domaine du traitement de l'information et de son exploitation.

Pour le bon fonctionnement de ce Master, un stage en entreprise est vivement recommandé dans le but de mettre en œuvre les différentes théories statistiques étudiées. C'est dans ce cadre que s'inscrit le présent stage que nous avons effectué dans la Compagnie d'Assurance **CHANAS ASSURANCES S.A**, agence de Yaoundé.

A l'issue de ce stage, nous présentons notre mémoire, résultat d'un travail de recherche effectué sous la supervision du Professeur Henri GWET, sur le thème : « **ANALYSE STATISTIQUE DU PROFIL DES CLIENTS A HAUT RISQUE DU PORTEFEUILLE AUTOMOBILE D'UNE COMPAGNIE D'ASSURANCE** »

Une telle étude s'avère nécessaire pour l'assureur afin de revoir sa tarification et surtout d'assurer la stabilité de ses provisions mathématiques.

Nous n'avons pas la prétention d'avoir cerné les contours du sujet, bien au contraire nous pensons que plusieurs études doivent encore être faites. Les données utilisées proviennent des services production et sinistre de la compagnie.

Résumé

Un préalable à la tarification des différents risques à assurer, est la connaissance a priori de ces risques qui constituent d'ailleurs la matière première dans l'industrie d'assurance.

La présente étude propose une méthode de détermination du profil de clients que l'on pourrait dans une certaine mesure qualifier de « client à risque » du portefeuille automobile de la Compagnie d'Assurance **CHANAS ASSURANCES S.A**, agence de Yaoundé.

Elle utilise pour ce faire l'analyse des correspondances multiples pour la description multidimensionnelle des données observées et le modèle de régression logistique pour la détermination des facteurs les plus pertinents et les plus discriminants expliquant de façon significative la sinistralité.

Outre les résultats d'analyse descriptive faite sur les données observées, il ressort principalement de cette étude que les **contrats temporaires** représentent un grand risque pour la compagnie ; et que les **véhicules de puissances fiscales comprises entre 11 et 14 chevaux(Essence) ou entre 8 et 10 chevaux (Diesel) sont les plus exposés au risque.**

Comme outil de classification automatique, un programme **R** a été proposé à l'assureur pour lui permettre d'affecter les nouveaux clients dans l'une des deux classes sans grand risque de se tromper.

Abstract

Before going to the tarification of the different risks assured, a knowledge prior to these risks, which constitute moreover the raw materials in the insurance industry appears necessary for promoters.

This study sets out to propose a determination method of client's profile, which to a certain extent could be described as the "client at risk" of the self-propelling portfolio of the **Insurance Company CHANAS ASSURANCES S.A, Yaoundé Branch.**

It uses in this case a multiple correspondence analysis for the multidimensional description of the data observed and the logistic regression method for the determination of most pertinent and discriminant factors, which can significantly explain the damage.

Besides the results of the descriptive analysis made on the observed data, this study reveals that **temporary contracts** represent a large risk for the company, and **vehicles with fiscal power between 11 and 14 horsepower (Petrol) or 8 and 10 horsepower (Diesel)** are the most exposed to risks.

As a tool for automatic classification, a program **R** has been proposed to the insurer to enable him to put new customers into one of the two categories without large risk of being deceived.

Chanass assurances s.a.

PRESENTATION DE CHANAS ASSURANCES S.A

CHANAS ASSURANCES S.A est une compagnie d'assurances **IARDT** (incendie, accident, risques divers, transport) régie par le code des assurances et présente au Cameroun depuis 1953 comme assureur et surtout comme facilitateur tant national qu'international dans le rapprochement des opérateurs connus de l'assurance que sont les assurés, les intermédiaires et les réassureurs.

Son siège social à Douala comporte plusieurs agences : Bafoussam, Yaoundé, Nkongsamba et une filiale en Guinée Equatorial (Malabo et Bata)
L'agence de Yaoundé qui nous a servi de structure d'accueil est située au boulevard Monseigneur Vogt, en face de la BICEC centrale.

Autrefois connu sous le nom de la **SARL CHANAS & PRIVAT ASSURANCES**, la compagnie est connue aujourd'hui sous la dénomination de **CHANAS ASSURANCES S.A** avec l'agrément ministériel du 24 Mars 2000. Sa progression et sa place de leader affirmé du marché camerounais des assurances n'est pas due au hasard mais reflète son sérieux et son professionnalisme. Ainsi, CHANAS se distingue particulièrement par :

- Son capital social, entièrement libéré de 2,3 milliards de FCFA représentant le plus important capital social des compagnies d'assurances des pays francophones de la sous-région, plus de quatre fois le minimum requis.
- Sa réassurance de premier ordre apéritée par l'un des leaders mondiaux, la MUNICH-RE
- Son chiffre d'affaire de plus de 16 milliards de FCFA en 2005 ; le plus élevé au Cameroun

CHANAS ASSURANCES S.A c'est aussi une grande expérience en matière de couverture des grands risques. Nous pouvons citer comme références :

Dans le domaine de l'aviation : CAMAIR

Dans l'offshore : la société nationale des hydrocarbures (SNH), TEXACO Cameroun, MOBIL OIL, COTCO

Dans les grandes unités de production : les Brasseries du Cameroun, le Chantier Naval et Industriel du Cameroun (CNIC), CRTV, AES-SONEL, les sociétés du groupe FOTSO (SAFCA, UNALOR, PILCAM, etc.)

Dans le tertiaire : Crédit lyonnais, CBC, les Ambassades (France, Russie, Espagne, Canada, Libye, Chine, Egypte)

Les organismes internationaux : HCR, UNICEF, OHADA, Centre Pasteur.

Dans les grands chantiers : EDOK-ETER, KETCH SCEMAR.

Pour ce qui est du domaine spécifique de l'assurance liée aux personnes, CHANAS ASSURANCES S.A a depuis longtemps développé un portefeuille MALADIE, qui lui confère aujourd'hui une expérience et une maîtrise notamment dans le remboursement des frais médicaux. Cette maîtrise lui permet aujourd'hui, d'offrir des garanties peu courantes dans tout le marché de l'assurance Camerounais. A titre d'exemple, on peut citer la couverture par les polices CHANAS ASSURANCES S.A de la pandémie du siècle, ce qui marque la détermination de la compagnie de participer avec l'état, à la lutte contre le SIDA.

INTRODUCTION

L'insécurité, intacte sur le fond et changeant dans sa forme ayant pris une ampleur, le développement des moyens de sécurisation humaine et matériel doit suivre le pas.

En effet, le besoin de sécurité naît naturellement de la prise de conscience des Hommes de se protéger contre les maladies, les accidents divers, et de façon générale contre l'aléa. Ceci a ouvert un segment à l'économie et à la gestion du risque où prospèrent prioritairement les compagnies d'assurance.

Les compagnies d'assurance sont des industries du secteur financier ayant pour but de soulager l'Homme dans la gestion des multiples risques liés à l'environnement dans lequel il évolue. L'assureur a donc pour tâche essentielle de transformer le risque qui constitue la matière première dans l'industrie d'assurance en plans d'assurance. Ainsi on pourrait voir les industries d'assurance comme étant des institutions de transformation de « l'incertitude » en « certitude ». Dès lors, l'assureur doit prendre des dispositions pour honorer ses promesses. C'est la raison pour laquelle il est nécessaire pour lui de faire des anticipations sur le devenir des différents risques assurés, et surtout de s'entourer d'un certain nombre d'indicateurs devant lui permettre de mieux jauger le risque avant toute tarification. La présente étude s'inscrit dans le cadre de participation à la solution de ce dernier.

Le thème soumis à notre attention est intitulé : « **Analyse statistique du profil des clients à haut risque du portefeuille automobile d'une compagnie d'assurance** » Etude que nous avons menée à la compagnie d'assurance CHANAS ASSURANCES S.A, agence de Yaoundé.

L'harmonisation et la sécurisation des différents risques routiers ont amené les gouvernements à imposer aux usagers la souscription d'une police d'assurance automobile. Ceci faisant de l'assurance automobile l'un des secteurs prioritaires du marché de l'assurance.

Dans l'industrie d'assurance, la tarification du risque fait l'objet des grandes théories du modèle stochastique de l'assurance. Dans la plupart des cas, on mène une étude approfondie des risques liés à chaque produit avant de faire une quelconque tarification. La question que l'assureur se pose est de savoir quelles sont les caractéristiques des clients que l'on pourrait qualifier de « haut sinistré ».

L'objectif de notre étude entre dans le cadre de réponse aux préoccupations des assureurs, puisqu'elle vise à **identifier les facteurs expliquant de façon significative la sinistralité ; ceci en vue de prévoir le statut (« haut risque » ou non) du nouveau client de la compagnie**. Pour atteindre cet objectif, nous avons structuré ce document en deux grandes parties :

La première fait un condensé des différentes théories Mathématiques et Statistiques mises en œuvre ;

La deuxième quant à elle est consacrée à la résolution du problème posé. Elle est constituée de trois chapitres :

Le **premier chapitre** décrit, code les variables étudiées et nous donne un bref aperçu sur le principe de fonctionnement du portefeuille automobile de la compagnie, de la souscription au règlement de sinistre.

Le **second chapitre** fait une description multidimensionnelle des données recueillies, à travers une analyse des correspondances multiples.

Le **troisième chapitre** quant à lui est consacré à la modélisation. Elle vise à construire le modèle s'ajustant le mieux aux données et ayant une bonne capacité de prédiction. Celui ci nous permettra non seulement de déterminer les variables les plus pertinentes et les plus discriminantes pouvant expliquer le risque automobile, mais aussi de proposer à l'assureur un algorithme permettant de dire au vu de ses caractéristiques si le nouveau client est à « haut risque » ou non.

Les données que nous analysons ici proviennent des services production et sinistre de la compagnie ; qui nous les a aimablement communiquées.

Chapitre Zéro

GENERALITES SUR L'ANALYSE DES CORRESPONDANCES MULTIPLES ET LA REGRESSION LOGISTIQUE

I - Analyse des correspondances multiples

A - Analyse générale

Considérons un tableau de valeurs numériques X ayant n lignes et p colonnes, correspondant à p variables recueillies sur n individus. On suppose $p < n$; et étant donnée une matrice M , M' désigne sa transposée.

On se propose de résoudre ici le problème de compression de données, c'est à dire de trouver les sous-espaces s'ajustant au mieux aux deux nuages de points (individus et variables). Pour exposer cette technique de réduction factorielle, nous nous plaçons successivement dans les espaces vectoriels R^p (des variables) et R^n (des individus)

a) Ajustements des nuages des individus et des variables

L'ajustement du nuage des individus se fait dans l'espace des variables et celui du nuage des variables dans l'espace des individus.

Chacune des n lignes du tableau X est considérée comme un point de R^p . De même chacune des p colonnes de X est considérée comme un point de R^n

Proposition 0-1

Le sous-espace à q dimensions ($q < p$) qui ajuste au mieux (au sens des moindres carrées) le nuage des points-individus dans R^p est engendré par les q premiers vecteurs propres de la matrice symétrique d'ordre (p,p) $X'X$ correspondant aux q plus grandes valeurs propres.

Le sous-espace à q dimensions qui ajuste au mieux le nuage des points-variables dans R^n est engendré par les q premiers vecteurs propres de la matrice symétrique d'ordre (n,n) XX' correspondant aux q plus grandes valeurs propres.

Toutes les valeurs propres non nulles des deux matrices $X'X$ et XX' sont égales.

Preuve [1]

Soit u_α le vecteur propre unitaire de $X'X$ correspondant à la valeur propre λ_α ; et v_α le vecteur propre unitaire de XX' correspondant à la même valeur propre λ_α .

Pour $\lambda_\alpha \neq 0$, on a les formules de transition entre les deux espaces R^p et R^n :

$$(1) \quad \begin{cases} v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X u_\alpha \\ u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X' v_\alpha \end{cases}$$

Dans R^p , u_α est le $\alpha^{\text{ième}}$ axe factoriel et l'on calcule le vecteur ψ_α des coordonnées sur cet axe par : $\psi_\alpha = X u_\alpha$

De même dans R^n , v_α est le $\alpha^{i\text{ème}}$ axe factoriel et l'on construit les coordonnées φ_α par :

$$\varphi_\alpha = X' v_\alpha$$

Compte tenu de (1), les facteurs peuvent se calculer par :

$$\begin{cases} \psi_\alpha = v_\alpha \sqrt{\lambda_\alpha} \\ \varphi_\alpha = u_\alpha \sqrt{\lambda_\alpha} \end{cases}$$

b) Diversification de l'analyse générale

Analyse générale avec des métriques et des critères quelconques

La métrique (la formule de distance) et le critère d'ajustement (pondération des points) varient suivant le problème et donc suivant la nature des variables.

Jusqu'à présent nous avons considéré les espaces munis de la matrice I (matrice identité) et nous avons supposé que tous les points du nuage avaient la même importance. Cependant il arrive que l'on ait à travailler avec une métrique plus générale et avec des individus dont les masses sont différentes. Généralisons le principe d'analyse factorielle présenté à des métriques et critères quelconques.

Plaçons-nous dans l'espace R^p et considérons le nuage de n points-lignes pesants. Soit X la matrice d'ordre (n, p) représentant le tableau des données, M la matrice symétrique définie positive d'ordre (p, p) définissant la métrique dans R^p , et N la matrice diagonale d'ordre (n, n) donc les éléments diagonaux sont les masses m_i des n points. Désignons par U la matrice d'ordre (p, p) ayant en colonne les vecteurs propres u_α (orthogonaux et unitaires) de XM .

Soit u un vecteur unitaire de R^p ($u'Mu = 1$) ; l'ensemble F des coordonnées des projections sur l'axe u des n points-lignes s'exprime par :

$$F = XMU$$

L'équation de l'axe factoriel u dans R^p s'écrit :

$X'NXMu = \lambda u$; et les coordonnées factorielles des n points sont données par la relation :
 $\psi = XMu$

Remarque 0-1

Si les masses et les métriques dans R^p (N et M) et dans R^n (P , matrice des masses des p points-colonnes et Q métrique dans R^n) n'ont pas de relation privilégiées entre elles, on perd les relations de transition.

En analyse des correspondances, on verra que la matrice des masses dans un espace est liée à la métrique de l'autre espace, ce qui permettra de conserver les relations de transition.

B - Analyse des correspondances

On considère ici deux variables qualitatives observées simultanément sur des individus. On suppose que la première variable notée X , possède n modalités notées $x_1, \dots, x_l, \dots, x_n$ et que la seconde, notée Y possède p modalités notées $y_1, \dots, y_h, \dots, y_p$.

Soit K le tableau de contingence à n lignes et p colonnes associé à ces observations.

A l'intersection d'une ligne i et d'une colonne j nous avons le nombre k_{ij} d'individus ayant simultanément les modalités x_i et y_j . Les totaux marginaux $k_{i.}$ et $k_{.j}$ représentent respectivement les nombres d'individus ayant la modalité x_i et y_j .

On a les relations suivantes :

$$k_{i.} = \sum_j k_{ij} ; k_{.j} = \sum_i k_{ij} ; k = \sum_{i,j} k_{ij}$$

Qui en terme de fréquences relatives donnent, lieu aux relations :

$$f_{ij} = \frac{k_{ij}}{k} ; f_{i.} = \sum_j f_{ij} ; f_{.j} = \sum_i f_{ij} ; \sum_{i,j} f_{ij} = 1$$

Pour analyser un tableau de contingence, on s'intéresse au tableau des profils-lignes et celui des profils-colonnes. Ainsi le $i^{\text{ième}}$ profil-ligne est : $\left\{ \frac{k_{i1}}{k_{i.}}, \dots, \frac{k_{ip}}{k_{i.}} \right\}$ et le $j^{\text{ième}}$ profil-colonne est :

$$\left\{ \frac{k_{1j}}{k_{.j}}, \dots, \frac{k_{nj}}{k_{.j}} \right\}$$

Distance du χ^2

Etant donnés deux profils-lignes i et i' on mesure leur écart à l'aide d'une métrique dite du χ^2 définie par :

$$d^2(i, i') = \sum_{j=1}^p \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

On définit de la même manière la distance entre les profils-colonnes.

a) Schéma général de l'analyse des correspondances

L'analyse des correspondances revient à effectuer l'analyse générale d'un nuage de points pondérés dans un espace muni de la métrique du χ^2 . On fera donc référence à l'analyse générale avec des métriques et des critères quelconques.

En analyse des correspondances, le tableau de données subit deux transformations, l'une en profils-lignes, l'autre en profils-colonnes, à partir desquelles vont être construits les nuages de points dans R^p et R^n .

Pour faire le lien avec l'analyse générale, nous conserverons les notations matricielles. Les transformations opérées sur le tableau des données peuvent s'écrire à partir des trois matrices F , D_n et D_p qui définissent les éléments de base de l'analyse.

F d'ordre (n, p) désigne le tableau des fréquences relatives ;

D_n d'ordre (n, n) est la matrice diagonale dont les éléments diagonaux sont les marges en lignes $f_{i.}$;

D_p est la matrice diagonale d'ordre (p, p) des marges en colonnes $f_{.j}$

Les deux nuages de points (dans l'espace des colonnes et dans l'espace des lignes) sont construits de manière analogue. Nous récapitulons ici les éléments de base de l'analyse qui vont permettre la construction des facteurs.

Tableau 0-1 : éléments de base de l'analyse des correspondances

Nuage de n points-lignes ← Dans l'espace R^p →	Eléments de base	Nuage de points-colonnes dans l'espace R^n
$X = D_n^{-1}F$ p coordonnées (point-ligne i) $\frac{f_{ij}}{f_i}$, pour $j = 1, 2, \dots, p$	Analyse du tableau X	$X = D_p^{-1}F'$ n coordonnées (point-colonne j) $\frac{f_{ij}}{f_j}$, pour $i = 1, 2, \dots, n$
$M = D_p^{-1}$ $d^2(i, i') = \sum_{j=1}^p \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2$	Avec la métrique M	$M = D_n^{-1}$ $d^2(j, j') = \sum_{i=1}^n \frac{1}{f_i} \left(\frac{f_{ij}}{f_j} - \frac{f_{i'j'}}{f_{j'}} \right)^2$
$N = D_n$ masse du point $i : f_i$	et le critère N	$N = D_p$ masse du point $j : f_{.j}$

b) Axes factoriels et facteurs

Nous supposons ici que p correspond à la plus petite dimension du tableau de données. Après avoir écarté la valeur propre triviale égale à 1 et le vecteur propre associé, nous retenons, de la diagonalisation de la matrice, les $p-1$ valeurs propres et les vecteurs propres associés. Nous obtenons ainsi au plus $p-1$ axes factoriels.

Tableau 0-2 : Eléments de construction de l'analyse des correspondances

← Dans R^p →	Eléments de construction	Dans R^n
$S = F'D_n^{-1}FD_p^{-1}$	Matrice à diagonaliser	$T = FD_p^{-1}F'D_n^{-1}$
$S u_\alpha = \lambda_\alpha u_\alpha$	Axe factoriel	$T v_\alpha = \lambda_\alpha v_\alpha$
$\psi_\alpha = D_n^{-1}FD_p^{-1}u_\alpha$ $\psi_{\alpha i} = \sum_{j=1}^p \frac{f_{ij}}{f_i \cdot f_{.j}} u_{\alpha j}$	Coordonnées factorielles	$\varphi_\alpha = D_p^{-1}F'D_n^{-1}v_\alpha$ $\varphi_{\alpha j} = \sum_{i=1}^n \frac{f_{ij}}{f_i \cdot f_{.j}} v_{\alpha i}$

C - Analyse des correspondances multiples

L'analyse des correspondances multiples (ACM) est une généralisation de l'analyse des correspondances, permettant de décrire les relations entre s ($s > 2$) variables qualitatives simultanément observées sur n individus.

On dispose ainsi d'un tableau de données R ayant n lignes et s colonnes mis sous forme de codage condensé. Le terme général r_{iq} désigne la modalité de la variable q choisie par le sujet i . En notant p_q le nombre de modalités de la variable q , on a $r_{iq} \leq p_q$. Mais un tel tableau n'est pas exploitable : les sommes en ligne et en colonnes n'ont pas de sens. Il faut donc recoder les variables.

a) Tableau disjonctif complet

Désignons par I l'ensemble des n individus et par p le nombre total des modalités des s variables. On a :
$$p = \sum_{q=1}^s p_q$$

On construit, à partir du tableau de données R , le tableau Z à n lignes et p colonnes décrivant les s variables pour les n individus par un codage binaire. Le tableau Z est la juxtaposition de s sous-tableaux : $Z = [Z_1, Z_2, \dots, Z_q, \dots, Z_s]$

Le sous-tableau Z_q à n lignes et p_q colonnes, est tel que sa $i^{\text{ème}}$ ligne contient $p_q - 1$ fois la valeur 0 et une fois la valeur 1 dans la colonne correspondant à la modalité de la variable q choisie par le sujet i . Autrement dit le tableau Z_q décrit la partition des n individus induite par les modalités de la variable q .

Le tableau Z est appelé tableau disjonctif complet dont le terme général s'écrit :

$z_{ij} = 1$ ou $z_{ij} = 0$ selon que le sujet i a choisi la modalité j de la variable q ou non.

Les marges en ligne du tableau disjonctif complet sont constantes et égales au nombre s de

variables :
$$z_{i.} = \sum_{j=1}^p z_{ij} = s$$

Les marges en colonnes $z_{.j} = \sum_{i=1}^n z_{ij}$ correspondent au nombre de sujets ayant choisi la modalité j de la variable q .

On vérifie que, pour chaque sous-tableau Z_q , l'effectif total est bien :
$$Z_q = \sum_{j \in q} z_{.j} = n$$

La somme des marges donne l'effectif total z du tableau Z soit :

$$z = \sum_{i=1}^n \sum_{j=1}^p z_{ij} = ns$$

b) Principes de l'analyse des correspondances multiples

L'analyse des correspondances multiples est l'analyse des correspondances d'un tableau disjonctif complet. Ses principes sont donc ceux de l'analyse des correspondances à savoir :

Même transformation du tableau de données en profils-lignes et en profils-colonnes ; même critère d'ajustement avec pondération des points par leurs profils marginaux ; même distance, celle du χ^2

c) Axes factoriels et facteurs

En prenant les résultats de l'analyse des correspondances et les notations adoptées, on pose :

$$F = \frac{1}{ns} Z \quad \text{de terme général} \quad f_{ij} = \frac{z_{ij}}{ns}; \quad D_p = \frac{1}{ns} D \quad \text{de terme générale} \quad f_{.j} = \delta_{ij} \frac{z_{.j}}{ns}$$

$$D_n = \frac{1}{n} I_n \quad \text{de terme générale} \quad f_{.i} = \frac{\delta_{ij}}{n}$$

où D est la matrice diagonale, d'ordre (p, p) d'effectif correspondant à chacune des modalités

des s variables ; I_n est la matrice identité d'ordre (n, n) et δ_{ij} tel que : $\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$

Pour trouver les axes factoriels u_α on diagonalise la matrice : $S = F'D_n^{-1}FD_p^{-1} = \frac{1}{s} Z'ZD^{-1}$

d) Inertie du nuage des modalités et conséquences pratiques

On rappelle que la distance du χ^2 dans R^n est la métrique D_n^{-1} . La distance entre la modalité j et le centre de gravité du nuage G , dont toutes les n coordonnées valent $\frac{1}{n}$, s'écrit :

$$d^2(j, G) = n \sum_{i=1}^n \left(\frac{z_{ij}}{z_{.j}} - \frac{1}{n} \right)^2 = \frac{n}{z_{.j}} - 1$$

L'inertie $I(j)$ de la modalité j vaut : $I(j) = m_j d^2(j, G)$; avec $m_j = \frac{z_{.j}}{ns}$; d'où :

$$I(j) = \frac{1}{s} \left(1 - \frac{z_{.j}}{n} \right)$$

On remarque que la part d'inertie due à une modalité de variable est d'autant plus grande que l'effectif dans cette modalité est plus faible. En conséquence, on évite, au moment du codage, les modalités à faibles effectifs susceptibles de perturber les directions des premiers axes factoriels.

L'inertie de la variable q , notée $I(q)$, vaut : $I(q) = \sum_{j=1}^{p_q} I(j) = \frac{1}{s} (p_q - 1)$

Ainsi la part d'inertie due à une variable est fonction croissante du nombre de modalités de la variable. D'où l'intérêt d'équilibrer le système des variables, c'est à dire le découpage des variables en modalités, si on veut faire jouer le même rôle à toutes les variables.

On en déduit que l'inertie totale du nuage des modalités vaut : $I = \sum_q I(q) = \frac{p}{s} - 1$

e) Règle d'interprétation

Compte tenu des distances entre les éléments du tableau disjonctif complet et des relations barycentriques particulières, on exprime :

La proximité entre modalités de variables différentes en terme d'association ;

La proximité entre deux modalités d'une même variable en terme de ressemblance

Deux séries de coefficients apportent une information supplémentaire par rapport aux coordonnées factorielles :

- Les contributions, parfois appelées contribution absolues, qui exprime la part prise par une modalité de la variable dans l'inertie (ou variance) « expliquée » par un facteur ;
- Les cosinus carrés, parfois appelés contributions relatives ou qualité de représentation, qui expriment la part prise par un facteur dans la dispersion d'une modalité de la variable .

C'est après l'examen de ces coefficients que l'on pourra interpréter les graphiques factoriels en tenant compte des relations de transition.

- Contributions

On cherche à connaître les éléments responsables de la construction de l'axe α .

On définit la contribution de l'élément j à l'axe α par : $Cr_{\alpha}(j) = \frac{f_{.j}\varphi_{\alpha j}^2}{\lambda_{\alpha}}$

Ce quotient permet de savoir dans quelle proportion un point j contribue à l'inertie λ_{α} du nuage projeté sur l'axe α .

Pour trouver une éventuelle signification à un axe, on s'intéresse d'abord aux points ayant une forte contribution. Ce sont eux qui fixent la position de l'axe (dans R^p pour les points i , et dans R^n pour les points j)

- Cosinus carrés

On cherche à apprécier si un point est bien représenté sur un sous-espace factoriel. Un point j dans R^n est plus ou moins proche de l'axe α .

La proximité entre deux points projetés sur l'axe α correspond d'autant mieux à leur distance réelle que les points sont plus proches de l'axe.

La « qualité » de la représentation du point j sur l'axe α peut être évaluée par le cosinus de l'angle entre l'axe joignant le centre de gravité du nuage au point j :

$$Cos_{\alpha}^2(j) = \frac{\varphi_{\alpha j}^2}{d^2(j, G)}$$

Plus le cosinus carré est proche de 1, plus la position du point observé en projection est proche de la position réelle du point dans l'espace. On apprécie la qualité de la représentation d'un point dans un plan en faisant la somme des cosinus carrés sur les axes étudiés.

Pour analyser les proximités entre points, on s'intéresse surtout aux points ayant un cosinus carré élevé. Les proximités entre ces points observés dans le sous-espace factoriel donnent une bonne image de leurs proximités réelles.

Remarque 0-2

Pour les contributions ainsi que pour les cosinus carrés, il n'y a pas de valeurs « seuils » à partir desquelles on peut dire que telle ou telle valeur est « forte » ou « faible ». Les appréciations se font empiriquement, en fonction de l'ensemble des valeurs calculées et varient d'un jeu de données à un autre. Cependant un critère de sélection des contributions significatives consiste à retenir les modalités de contribution supérieure au poids [3]

- Nombre d'axes à retenir

On préconise généralement de détecter sur le diagramme des valeurs propres l'existence d'un coude, ce qui n'est pas toujours aisé en pratique. Le scree-test de Cattell [3] en est la version analytique. On calcule les différences premières (des valeurs propres décroissantes) :

$$\lambda_1 - \lambda_2 = \varepsilon_1 \quad ; \quad \lambda_2 - \lambda_3 = \varepsilon_2 \quad \dots$$

Puis les différences secondes : $\varepsilon_1 - \varepsilon_2 = \delta_1$; $\varepsilon_2 - \varepsilon_3 = \delta_2 \dots$

On retient alors les valeurs propres $\lambda_1, \lambda_2, \dots, \lambda_k, \lambda_{k+1}$ telles que $\delta_1, \delta_2, \dots, \delta_k$ soient tous positifs.

f) Variables supplémentaires

L'utilisation des variables supplémentaires en analyse des correspondances multiples permet d'enrichir l'interprétation des axes par des variables n'ayant pas participé à leur construction. L. Lebart et A. Morineau [1] ont introduit la notion de valeur-test pour chaque modalité d'une variable afin de juger si le point représentatif de la modalité est significativement différent de la moyenne générale.

Le principe en est le suivant. Pour évaluer l'ampleur des différences entre proportions ou entre moyennes, on réalise des tests statistiques que l'on exprime finalement en nombre d'écart types d'une loi normale. La valeur-test est égale à ce nombre d'écart types.

Ainsi lorsque la valeur-test est supérieure à 2 en valeur absolue, un écart est significatif au seuil usuel (5%).

II – Régression Logistique

A – Introduction

On étudie la modélisation de données endogènes binaires Y à partir de conditions exogènes x . L'endogène Y est par exemple l'état de santé d'un individu (sain ou malade), le statut d'un sinistré (« haut sinistré » ou non).

Nous nous intéresserons ici au cas où la variable exogène x à état dans E est qualitative. Codons par 0 et 1 les deux états de l'endogène Y . Y sous la condition x est une variable de Bernoulli $B(\pi(x))$ caractérisée par la probabilité :

$$\Pr(Y = 1/x) = \pi(x)$$

On a : $y = E(Y/x) + \varepsilon$, avec $E(Y/x) = P_r(Y = 1/x)$;

$E(\cdot)$ étant l'espérance mathématique et ε le résidu associé à y

Pour fixer les idées, supposons que l'espace d'état de x est $E = \mathbb{R}^p$.

Soit $F : \mathbb{R} \rightarrow [0,1]$ une fonction continue.

Une façon de modéliser $\pi(\cdot)$ est d'écrire :

$$\pi(x) = F(\beta'x)$$

où $\beta \in \mathbb{R}^p$ est un paramètre inconnu. Généralement on choisit pour F une fonction de répartition (noté fdr). Le modèle est linéaire en β à travers F , une fonction non linéaire. F est la fonction de lien du modèle (link function).

Codage d'une variable exogène qualitative

Si une composante z de x est qualitative, $z \in \{a_1, a_2, \dots, a_k\}$, z peut être codé dans \mathbb{R}^{k-1} en identifiant a_l au $l^{\text{ième}}$ vecteur de la base canonique de \mathbb{R}^{k-1} , $l = 1, k-1$, et a_k égal à 0.

B - Le modèle logit

La distribution logistique est associée à la fdr $\Lambda(u) = \frac{e^u}{1+e^u}$.

La fonction $Logit : [0,1] \rightarrow \bar{R}$ est la fonction réciproque de Λ ,

$$Logit(y) = \log \frac{y}{1-y}$$

Le modèle $Logit$ est défini par :

$$\pi(x) = \Pr(Y = 1/x) = \Lambda(' \beta x) \quad \text{ou} \quad Logit(\pi(x)) = ' \beta x$$

C – Estimations et tests

a) Estimation du modèle $Logit$

Pour l'estimation des coefficients d'un modèle de régression logistique, la méthode généralement utilisée est celle du maximum de vraisemblance (M.V). On peut décrire sommairement cette méthode comme suit :

Soit Y une variable qui obéit à une loi de distribution de paramètre $\beta : f(Y; \beta)$. A partir d'un certain nombre d'observations sur Y , (Y_1, Y_2, \dots, Y_n) , on essaie de déterminer la valeur inconnue du paramètre β . La méthode du maximum de vraisemblance postule que cette valeur de β devrait être celle qui maximise la probabilité d'obtenir les valeurs observées sur Y . Lorsque les observations individuelles $y_i, i = 1, \dots, n$ sont supposées indépendantes, cette vraisemblance s'écrit comme le produit des probabilités :

$$L(\beta) = \prod_{i=1}^n [\pi(x, \beta)]^{y_i} [1 - \pi(x, \beta)]^{(1-y_i)}$$

Ensuite, on maximise cette vraisemblance par rapport au paramètre β au moyen d'un Algorithme numérique (par exemple une méthode de gradient).

b) Test de sous-modèle

Le test de sous-modèle est basé sur le test du rapport de vraisemblance [2]

Si $M_1 \subset M_2$ sont deux modèles emboîtés de dimensions $p_1 < p_2$;

Si L_n est la log-vraisemblance calculée à la valeur $\hat{\beta}_n(M)$, l'estimateur du maximum de vraisemblance de β sous M , alors, sous M_1 :

$$2\{L_n(M_2) - L_n(M_1)\} \xrightarrow{\text{loi}} \chi^2(p_2 - p_1) \quad (\text{Chi deux à } p_2 - p_1 \text{ degré(s) de liberté})$$

c) Critère de choix de modèle de type AIC

En régression logistique, l'un des critères de choix du modèle à retenir est l'AIC (Akaike Information Criterium) : Le modèle étant d'autant plus intéressant que son AIC est faible.

Pour un modèle M donné, on a :

$$AIC(M) = -2(\log L_n(M) + d(M)) ; \quad d(M) \text{ étant la dimension de } M .$$

Test du Chi deux (χ^2) d'indépendance

Le test de χ^2 d'indépendance s'applique à l'étude de la liaison entre deux variables qualitatives X et Y . On teste les hypothèses :

Hypothèse nulle (H_0) : les deux variables sont indépendantes

Contre

Hypothèse alternative (H_1) : les deux variables sont liées

Principe du test

On observe un n-échantillon du couple (X, Y) de variables aléatoires X à valeurs $\{1, \dots, k\}$ et Y à valeurs $\{1, \dots, l\}$.

Soit $p = \{p_{ij} : 1 \leq i \leq k, 1 \leq j \leq l\}$ la loi de (X, Y) .

Si N_{ij} est le nombre d'observation de (i, j) dans le n-échantillon, l'estimateur empirique \overline{p}_n est défini par :

$$\overline{p}_n(i, j) = \frac{N_{ij}}{n}$$

Si les caractères X et Y sont indépendants, p est dans l'ensemble des lois produits :

$\theta = \{p : p_{ij} = p_{i.} p_{.j} \text{ pour } 1 \leq i \leq k, 1 \leq j \leq l\}$ (le point représente la sommation sur l'indice)

On estime alors $p_{i.}$ par l'estimateur empirique $\frac{N_{i.}}{n}$ et $p_{.j}$ par $\frac{N_{.j}}{n}$;

Donc p_{ij} par $\frac{N_{i.} N_{.j}}{n^2} = \hat{p}_n(i, j)$

Si H_0 est vraie, \hat{p}_n et \overline{p}_n doivent être voisins ; le χ^2 d'indépendance

$$\chi_n^2(\hat{p}_n, \overline{p}_n) = n \sum_{i,j} \frac{\left(\frac{N_{ij}}{n} - \frac{N_{i.} N_{.j}}{n^2} \right)^2}{\frac{N_{i.} N_{.j}}{n^2}}$$

ne doit pas être très grand.

On montre [10] que $\chi_n^2(\hat{p}_n, \overline{p}_n)$ converge en loi vers $\chi^2((k-1)(l-1))$ (loi du Chi deux à $(k-1)(l-1)$ degrés de libertés)

Un test convenable pour tester l'indépendance de X et Y avec un niveau voisin de α est le test de région de rejet (de l'hypothèse nulle)

$$R_\alpha = \left\{ \chi_n^2(\hat{p}_n, \overline{p}_n) > \chi^2((k-1)(l-1), \alpha) \right\}$$

NB : Pour un test d'hypothèses H_0 (hypothèse nulle) contre H_1 (hypothèse alternative), on appelle **p-value** la probabilité sous l'hypothèse nulle, que la statistique observée soit supérieure à la statistique théorique. Dans le cas du test du Chi deux ci dessus, on a :

$$\text{p-value} = \Pr_{H_0} (\chi_n^2(\hat{P}_n, \overline{P}_n) \in R_\alpha)$$

Si $\text{p-value} \leq \alpha$, on dit que le test est significatif à $\alpha\%$.

CHAPITRE PREMIER

DESCRIPTION DES VARIABLES ET CODAGE

Les données dont nous disposons dans cette étude proviennent des services production et sinistre de la compagnie CHANAS ASSURANCES S.A, agence de Yaoundé. Où nous avons comme population d'étude les sinistrés du portefeuille automobile pendant une période donnée (premier Janvier 2005 au 13 juin 2005).

Avant de présenter les données de façon explicite, nous donnerons un bref aperçu sur le principe de fonctionnement de l'assurance automobile de la compagnie.

I - Principe de fonctionnement de l'assurance automobile : de la souscription au règlement de sinistre

Lors de la souscription d'un contrat d'assurance automobile au service production de la compagnie, un document décrivant les différentes clauses possibles du contrat est remis au souscripteur. Puis, il lui est remis un questionnaire à remplir ; comportant les renseignements sur le véhicule à assurer, le conducteur du véhicule, les différentes garanties à souscrire et la durée du contrat.

- Véhicule à assurer

Au moyen de la carte grise du véhicule, on extrait les informations telles que : la marque, le genre, la puissance fiscale, son âge (à partir de sa date de première mise en circulation), et le souscripteur donne l'usage qu'il fera du véhicule.

- Différentes garanties

Entre autres, nous avons les garanties : responsabilité civile, bris de glace, dommages au véhicule, vols total et partiel, honoraires d'expert, recours défense, recours tiers incendie, incendie, individuelle personnes transportées, braquage, etc.

- Conducteur du véhicule

Ici on a les informations telles que : son adresse, sa profession, son âge, son coefficient de bonification.

Toutes ces informations conduisent à la fixation d'une prime (prime émise) à payer par le souscripteur de contrat afin d'être couvert en cas de sinistre durant la période de garantie.

Rappelons aussi que les polices d'assurance automobile sont classées en deux grandes catégories :

Les polices mono, constituées d'un seul véhicule.

Les polices flottes, constituées de plusieurs véhicules appartenant le plus souvent à un groupe de personnes exerçant la même activité, ou à un particulier.

Pour satisfaire ses clients, la compagnie octroie des bonus aux polices flottes, ceci en fonction du nombre de véhicule de la flotte. Il est cependant important de noter que la bonification accordée aux polices flottes n'a rien à voir avec le coefficient de bonification individuelle qui tient compte de la responsabilité civile et de l'ancienneté du client dans la compagnie. Ainsi, avec une attestation de non sinistre, le client qu'il soit nouveau ou ancien peut se retrouver avec un bonus pour non sinistre (BNS) allant jusqu'à 25 % de la prime émise.

Dans notre étude, afin de comparer les clients sur des bases communes, nous n'avons pas jugé utile de considérer la bonification liée aux polices flottes. Ceci étant, nous ne considérons ici que le bonus individuel (lié à chaque conducteur), qui peut nous servir comme outil de mesure de responsabilité civile du client. A priori on pourra donc qualifier de bon conducteur celui dont le coefficient de bonification est maximal.

Une fois la souscription faite, après la survenance d'un sinistre, l'assuré se présente au service sinistre où il remplit une fiche de déclaration de sinistre dans laquelle on retrouve les caractéristiques du véhicule sinistré et du conducteur du véhicule au moment du sinistre. On peut avoir recours à un expert qui examine le sinistre et arrête son coût financier qui est fonction des différentes garanties souscrites.

II -Méthodologie de collecte de données

La collecte des données a été la phase la plus difficile de ce travail ; ceci à cause du fait qu'elle soit effectuée manuellement.

A partir du progiciel EXTEL géré sous l'AS400, nous avons extrait les numéros de police automobile ayant eu la réalisation d'au moins un sinistre, du premier janvier 2005 au 13 juin, date de début de notre étude.

Après avoir repéré les numéros de police comportant les véhicules sinistrés, nous avons eu recours au service production où il était question pour nous, compte tenu du fait que notre unité statistique soit le véhicule sinistré, d'identifier exactement le(s) véhicule(s) sinistré(s) dans le cas des polices flottes. L'identification se faisant à partir de la marque du véhicule sinistré et de son numéro d'immatriculation, recueillis au service sinistre de la compagnie.

Une fois l'identification faite, outre les différents coûts de sinistres, on est passé au recueil des différentes informations fournies lors de la souscription du contrat.

Compte tenu de l'objectif de notre sujet, il convient de mentionner que nous nous intéresserons particulièrement aux différents coûts de sinistres. Peu importe leurs nombres. Ainsi pour un véhicule voyant plus d'une fois la survenance d'un sinistre pendant la même période de garantie, on s'intéressera uniquement à la somme de leurs différents coûts.

Les données recueillies ont été saisies dans l'éditeur du logiciel statistique SPSS, pour leur apurement. Apurement qui consistait principalement à supprimer les doublons et aberrations dues au recueil manuel de données.

Après l'apurement par SPSS de la base de données, nous avons obtenu une nouvelle base comportant 229 enregistrements correspondant à notre population de sinistrés, parmi lesquels nous retrouverons 90 que nous qualifierons de « hauts sinistrés ». Cette notion sera définie dans la suite.

A) Définition de « haut sinistré »

Habituellement, un sinistre est déclaré « *haut sinistre* » lorsque son coût représente au moins 70 % de la prime émise. Mais il se trouve que dans notre échantillon, la quasi totalité des sinistrés vérifient ce critère. Ce qui rendrait peu intéressante la variable sinistralité. Pour remédier à cette situation, nous avons défini une nouvelle règle de classification à savoir :

- Pour les contrats arrivés à échéance avant le début de l'étude :

Si le coût du sinistre est supérieur ou égal à quatre fois la prime émise, le sinistré sera déclaré « haut sinistré ».

- Pour les contrats en cours au moment de l'étude :

On calcule une nouvelle durée de contrat ; qui est le temps en mois séparant la souscription du dit contrat et le début de l'étude (13 juin 2005). Dans ce cas le critère de discrimination tient compte des deux durées de contrat ; étant donné que la prime émise a été fixée en fonction de la durée initiale du contrat. Ainsi, si :

Le coût du sinistre est supérieur ou égal à quatre fois la nouvelle durée du contrat multiplié par la prime émise divisé par la durée initiale du contrat, le sinistré sera qualifié « haut sinistré ».

B) Description et codage des variables

Les différentes variables utilisées dans cette étude peuvent être scindées en deux groupes, à savoir : la variable à expliquer ou variable endogène, et les variables explicatives ou variables exogènes.

a) Variable à expliquer

Nous avons ici la variable réponse correspondante à l'objectif visé par notre étude.

SINISTRALITE : variable dichotomique prenant la valeur :

- 1** si le sinistre est qualifié de « haut sinistré » tel que défini ci haut
- 0** sinon

b) Variables explicatives

Ce sont pour la plupart les variables décrivant le risque automobile. Elles peuvent être classées en trois sous groupes : les variables représentant les garanties souscrites, les variables décrivant les caractéristiques du véhicule et les variables représentant le niveau de prime, durée de contrat et coût de sinistre.

- Variables représentant les garanties souscrites

Ce sont des variables booléennes prenant les valeurs 0 et 1.

- 1** si la garantie a été souscrite et
- 0** sinon.

Les principales sont les suivantes :

BG : bris de glace

DOM : dommage véhicule

VOLX : vol partiel du véhicule

VOL : vol total du véhicule

HE : honoraire d'expert

INC : incendie totale

INCX : incendie partiel
IPT : individuelle personne transportée
RD : recours défense
RTI : recours tiers incendie
MI : matière inflammable
BRAQ : braquage

Rappelons ici que la garantie VOL est incluse dans la garantie VOLX, et compte tenu de l'effectif très réduit de souscription de garantie VOLX par rapport à la garantie VOL, nous avons réuni les deux pour en faire la garantie VOL. Il en est de même pour les garanties INC et INCX.

Les garanties MI et RTI ne sont pas prises en compte dans cette étude. Ceci pour éviter le problème de sous dispersion des modalités 1 ; car la quasi totalité des véhicules de notre échantillon n'ont pas souscrit à ces garanties. Elles représentent respectivement 1,17 % et 1,95 % de l'échantillon global.

- Variable représentant les caractéristiques du véhicule

MARQUE : Variable qualitative, représentant la marque du véhicule.

Nous avons regroupé les différentes marques en quatre niveaux, les trois premiers représentant les pays d'origine des véhicules. Nous avons :

Japon : Représentant les véhicules de marque Toyota, Mitsubishi, Suzuki, Isuzu, Mazda, Subaru, Yamaha, Honda, nissan.

France : Représentant les véhicules de marque Peugeot, Renault, Citroën.

Allemagne : Représentant les véhicules de marque Golf, BMW, Volkswagen, Audi, mercedes.

Autres : Pour les autres marques de véhicules.

GENRE : Variable qualitative représentant le genre de véhicule.

Conjointement avec le service production de la compagnie, nous avons classé les valeurs de cette variable en trois catégories, suivant la configuration des différents véhicules. Ainsi, on a :

Genre1 : Vélo, cyclomoteur, moto, solo, berline, pajero, sw (court chassis), vp (voiture personnelle), ci

Genre2 : Fourgon, pu (pick-up), ctte (camionnette), bâchée, bus, autocar

Genre3 : Semi remorque, camion, caterpillar, tracteur

USAGE : Variable qualitative, représentant l'usage que le souscripteur fera de son véhicule.

Pour ce qui est des définitions de différents usages de véhicules, il convient de se référer au titre1 « *Déclarations relatives à l'usage du véhicule assuré* » de la nomenclature des clauses de l'assurance automobile.

PUISS : Variable qualitative ordinale à six niveaux, représentant la puissance fiscale du véhicule.

Nous avons dans notre échantillon des véhicules à moteur diesel et des véhicules à essence. Malgré ces différentes sources d'énergie, rappelons ici qu'il ne s'est pas posé le problème d'équivalence de puissance. Ceci à l'aide du document d'équivalence de puissance fiscale, fourni par la société camerounaise d'assurance ; document que nous prendrons le soin de joindre en annexe. Ainsi la variable PUISS aura six niveaux : PUISS1, ..., PUISS6, suivant les valeurs croissantes de puissances fiscales.

- Variables représentant les niveaux de primes, durée de contrat et coût de sinistre.

PRIME : Variable quantitative discrète représentant la prime émise par le souscripteur pour être couvert par l'assureur en cas de sinistre pendant la validité du contrat.

COEFB : Variable qualitative ordinale à cinq niveaux : 0%, 10%, 15%, 20%, 25% représentant le coefficient de bonification du souscripteur.

Compte tenu des critères d'attribution de bonus définis plus haut, on pourra à priori l'utiliser comme outil de mesure de responsabilité civile chez les conducteurs. Dès lors, le bon conducteur aura le plus grand coefficient de bonification.

DURC : Variable quantitative discrète, représentant la durée de contrat en mois. Celle-ci étant comprise entre 1 et 12 mois. Le contrat sera dit temporaire si sa durée est inférieure à 12 mois.

COUTSIN : Variable quantitative discrète représentant le coût du sinistre, c'est à dire le montant déboursé par la compagnie pour couvrir le sinistre.

$$x_1, \dots, x_l, \dots, x_n$$

CHAPITRE DEUX

DESCRIPTION MULTIDIMENSIONNELLE DES DONNEES

Dans ce chapitre, nous avons pour souci de décrire les rapprochements pouvant exister entre les différentes modalités de notre base de données et de spécifier celles qui se rapprochent le plus de la caractéristique « haute sinistralité » (SIN=1). Pour cela nous utiliserons une méthode classique de description multidimensionnelle de données qualitatives : L'analyse des correspondances multiples (ACM).

I - Analyse des correspondances multiples

Définition

L'analyse des correspondances multiples est une méthode d'analyse de données qui consiste à décrire les proximités entre les modalités de variables qualitatives simultanément observées sur des individus. Elle est aussi utilisée pour la construction de scores comme préalable à une méthode de classification nécessitant les données qualitatives.

A - Condition d'utilisation de la méthode

Comme toute méthode d'analyse de données, l'ACM s'applique après vérification d'un certain nombre de critères conduisant à sa robustesse. Nous entendons par là :

- Le respect de la non disparité trop grande entre les nombres de catégories des variables.
- Le respect de l'écart pas trop grand entre les fréquences des modalités d'une même variable.

Nous utilisons pour ce chapitre le logiciel statistique SPAD version 4.01. Avant de passer à l'analyse, il convient pour nous de vérifier si la base de données brutes respecte les critères d'utilisation de l'ACM mentionnés ci-dessus. Pour cela, le **tableau 2-1** ci-dessous nous donne les statistiques élémentaires des différentes variables étudiées.

Tableau 2-1 : Statistiques élémentaires des données brutes

	----- EFFECTIFS -----		
	ABSOLU	%/TOTAL	HISTOGRAMME DES POIDS
SIN			
0	139	60.70	*****
1	90	39.30	*****
	229	100.00	
BG			
0	214	93.45	*****
1	15	6.55	****
	229	100.00	
DOM			
0	206	89.96	*****
1	23	10.04	*****
	229	100.00	
VOL			
0	106	46.29	*****
1	123	53.71	*****
	229	100.00	
HE			
0	144	62.88	*****
1	85	37.12	*****
	229	100.00	
INC			
0	108	47.16	*****
1	121	52.84	*****
	229	100.00	
IPT			
0	16	6.99	****
1	213	93.01	*****
	229	100.00	
RD			
0	40	17.47	*****
1	189	82.53	*****
	229	100.00	
BRAQ			
0	178	77.73	*****
1	51	22.27	*****
	229	100.00	
MARQUE			
all	37	16.16	*****
autre	26	11.35	*****
france	42	18.34	*****
japon	124	54.15	*****
	229	100.00	
GENRE			
1	164	71.62	*****
2	31	13.54	*****
3	34	14.85	*****
	229	100.00	
USAGE			
1	156	68.12	*****
10	1	0.44	*
2	45	19.65	*****
3	16	6.99	****
4	6	2.62	**
5	3	1.31	*
8	1	0.44	*
9	1	0.44	*
	229	100.00	

COEFB			
0	123	53.71	*****
10	24	10.48	*****
15	10	4.37	***
20	18	7.86	****
25	54	23.58	*****
	229	100.00	

PUISS			
1	3	1.31	*
2	1	0.44	*
3	75	32.75	*****
4	77	33.62	*****
5	37	16.16	*****
6	36	15.72	*****
	229	100.00	

DURC			
10	8	3.49	**
12	177	77.29	*****
2	12	5.24	***
3	11	4.80	***
4	3	1.31	*
5	1	0.44	*
6	8	3.49	**
7	2	0.87	*
8	3	1.31	*
9	4	1.75	*
	229	100.00	

Commentaire 2-1

L'histogramme des poids de modalités présenté dans le tableau ci-dessus nous permet de lire d'une part un grand déséquilibre entre les poids des modalités de certaines variables, d'autre part une grande disparité de nombre de modalités des différentes variables. Ce problème de déséquilibre de nombre de modalités se voit aisément lorsqu'on considère par exemple les variables GENRE (trois niveaux) et DURC (dix niveaux).

On constate donc que pour appliquer la méthode d'analyse des correspondances multiples, notre base de données a besoin d'être recodée, ceci en vue d'être le plus proche possible des critères d'application de la méthode.

B - Recodage de variables

Il se fait dans le souci de satisfaction des conditions d'application de la méthode d'analyse des correspondances multiples. Nous recodons ainsi une bonne partie de nos variables, notamment :

COEFB : Variable qualitative ordinaire à 3 modalités, représentant le coefficient de bonification du client. Les différentes modalités sont :

- 0 pour les clients n'ayant pas de bonus
- 1 pour les clients ayant un bonus de 10 % ou 15 %
- 2 pour les clients ayant un bonus de 20 % ou 25 %

PUISS : Variable qualitative ordinaire à 3 modalités, représentant la puissance fiscale du véhicule. Ses modalités sont :

- 1 pour les véhicules à essence de puissance inférieure ou égale à 10 chevaux, ou les véhicules diesel de puissance inférieure ou égale à 7 chevaux.
- 2 pour les véhicules à essence de puissances comprises entre 11 et 14 chevaux, ou les véhicules diesel de puissances comprises entre 8 et 10 chevaux.
- 3 pour les véhicules à essence de puissances supérieures à 15 chevaux, ou les véhicules diesel de puissances supérieures à 11 chevaux.

USAGE : Variable qualitative à 4 modalités, représentant l'usage du véhicule. Ses modalités sont :

- 1 : Véhicules utilisés pour l'exercice d'une profession et pour la promenade.
- 2 : Véhicules utilisés pour le transport des produits ou marchandises appartenant à l'assuré
- 3 : Véhicules utilisés à des transports à titre onéreux de produits ou marchandises appartenant à des tiers
- 4 : 50 % (Autobus, autocar et véhicules pour le transport voyageur à titre payant.) ;
25 % (Véhicule motorisés à deux roues) ;
8,33 % (Véhicules destinés à la location sans chauffeur) ;
8,33 % (Véhicules utilisés par des entreprises industrielles pour l'exécution des travaux de chantier) ;
8,33 % (Ambulances, corbillards, Fourgon funéraires).

DURC : Variable qualitative dichotomique représentant le type de contrat. Ses modalités sont :

- a : pour les contrats annuels, c'est à dire de durée égale à 12 mois ;
- t : pour les contrats temporaires, c'est à dire de durée inférieure à 12 mois.

Après avoir recodé les variables, nous obtenons une nouvelle base de données non loin des exigences que relève l'analyse des correspondances multiples. Cette nouvelle base est à nouveau transférée dans le logiciel pour l'analyse.

C - Principe de l'analyse

Rappelons une fois de plus que l'analyse des correspondances multiples utilisée ici a pour principal objectif de déterminer les rapprochements existant entre les différentes modalités de la base de données ; rapprochement pouvant s'interpréter comme une association entre ces modalités. Dès lors nous pouvons formuler les hypothèses de corrélation entre les variables correspondantes aux modalités associées.

Dans cette analyse, les variables décrivant les différentes garanties souscrites seront prises comme supplémentaires, c'est à dire ne participeront pas à la construction des axes factoriels. Ceci se fait principalement dans le but d'enrichir l'interprétation des axes par des variables n'ayant pas participé à leurs déterminations.

L'interprétation des listings d'une ACM est principalement basée sur l'examen :

- Du cosinus carrés de l'angle entre l'axe factoriel et la droite reliant un point-modalité au centre de gravité du nuage de points ;
- Des contributions des modalités à la construction des axes factoriels ;
- Des coordonnées des modalités dans les différents axes factoriels

Nous commençons par regarder les modalités bien représentées sur les axes, c'est à dire celles ayant un bon cosinus carré. Ensuite, on admettra une modalité comme significativement contributive à la construction d'un axe factoriel si sa contribution est supérieure à son poids [3]. Enfin nous nous servirons des coordonnées des modalités pour mieux orienter celles qui seront retenues pour un axe donné.

Pour ce qui est des variables supplémentaires, nous utilisons la notion des valeurs-tests [1]. Une modalité de variable supplémentaire sera attachée à un axe si sa valeur-test associée à l'axe est en valeur absolue supérieure à 2.

Pour mieux concrétiser les interprétations que nous ferons, nous proposerons les graphiques représentant les projections du nuage des modalités dans les plans principaux. Cependant nous nous garderons des illusions de proximités pour les modalités qui ne sont pas bien représentées dans un plan donné.

D - Interprétations

Le scree-test de Cattell [3] nous recommande le choix des trois axes correspondant aux trois premières valeurs propres. Cependant au vu de l'histogramme des valeurs propres présenté au **tableau 2-2**, nous pouvons compte tenu du décrochement[4] considérable entre la quatrième et la cinquième valeur propre, faire une description du nuage avec les quatre premiers axes. L'étalement figurant sur ces quatre axes représente 50,77% de l'inertie initiale du nuage des modalités.

Comme nous l'avons dit plus haut, l'interprétation des axes par les variables actives se fait essentiellement à l'aide des trois entités : cosinus carrés, contributions et coordonnées. Ainsi le **tableau 2-3** nous donne les cosinus carrés, les contributions et les coordonnées des différentes modalités dans les axes factoriels. Nous avons mis en gras les valeurs correspondantes aux modalités bien représentées (bon cosinus carré) et celles correspondantes aux modalités de bonne contribution à l'inertie de l'axe (contribution supérieure au poids). Le **tableau 2-4** nous donne un récapitulatif des modalités bien représentées et de bonnes contributions pour chacun des quatre axes retenus. L'interprétation des axes par les variables supplémentaires est donnée par le **tableau 2-5**.

a) Interprétation des axes factoriels

L'axe 1 oppose :

D'une part les véhicules de **genre 3** (camion, semi remorque, caterpillar, tracteur), **d'usage 3** (utilisés pour des transports à titre onéreux des produits ou marchandises appartenant à des tiers), et de **puissance 3** (puissance fiscale supérieure à 15 chevaux pour essence et 11 chevaux pour diesel) ;

D'autre part les véhicules de **genre1** (petits véhicules) et **d'usage 1** (utilisés pour l'exercice d'une profession et pour la promenade) qui sont particulièrement caractérisés par la souscription aux garanties *vol, braquage et incendie*

L'axe 2 oppose :

D'une part les véhicules de **genre 2** (Fourgons, bachées, bus, autocars...), de **puissance 2** (puissances fiscales comprises entre 11 et 14 chevaux pour essence et entre 8 et 10 chevaux pour diesel), **d'usage 2 ou 4** (utilisés pour le transport de produits ou marchandises appartenant à l'assuré ou d'usage 4(cf recodage des variables)) et la modalité « **haute sinistralité** »;

D'autre part les véhicules **d'usage 3**, caractérisés par la souscription à la garantie *incendie* et la modalité « **non haute sinistralité** »

L'axe 3 oppose :

D'une part les **contrats temporaires**, caractérisés par la non souscription aux garanties *vol* et *incendie* et principalement par la mention « **haute sinistralité** » ;

D'autre part les **contrats annuels** réalisés pour des véhicules **d'usage 2**

L'axe 4 oppose :

Les véhicules **d'origine allemande**, de **puissances 2** aux véhicules **d'origine** « **autres** »(ni française, ni japonaise), essentiellement souscripteurs aux garanties *vol* et *incendie*.

Tableau 2-2 :Histogramme des valeurs propres

NUMERO	VALEUR	POURCENT.	POURCENT.	HISTOGRAMME DES 14 PREMIERES VALEURS PROPRES
	PROPRE		CUMULE	
1	0.3552	17.76	17.76	*****
2	0.2711	13.55	31.31	*****
3	0.1995	9.98	41.29	*****
4	0.1897	9.48	50.77	*****
5	0.1543	7.72	58.49	*****
6	0.1404	7.02	65.51	*****
7	0.1320	6.60	72.11	*****
8	0.1276	6.38	78.49	*****
9	0.1094	5.47	83.96	*****
10	0.1004	5.02	88.98	*****
11	0.0964	4.82	93.80	*****
12	0.0687	3.44	97.24	*****
13	0.0466	2.33	99.57	*****
14	0.0087	0.43	100.00	**

Tableau 2-3 : Coordonnées, Contributions et cosinus carrés des modalités actives sur les axes 1 à 4

MODALITES			COORDONNEES				CONTRIBUTIONS				COSINUS CARRÉS			
LIBELLE	P.REL	DISTO	1	2	3	4	1	2	3	4	1	2	3	4
SIN														
0	8.67	0.65	-0.05	0.31	-0.47	-0.10	0.1	3.1	9.4	0.4	0.00	0.15	0.33	0.01
1	5.61	1.54	0.07	-0.48	0.72	0.15	0.1	4.7	14.5	0.6	0.00	0.15	0.33	0.01
---CONTRIBUTION CUMULEE =							0.1	7.8	23.9	1.1	+-----			
MARQUE														
all	2.31	5.19	0.51	0.03	0.40	-1.52	1.7	0.0	1.8	28.3	0.05	0.00	0.03	0.45
autre	1.62	7.81	0.92	0.03	0.80	1.29	3.9	0.0	5.2	14.1	0.11	0.00	0.08	0.21
france	2.62	4.45	-0.41	0.76	-0.53	0.57	1.3	5.6	3.7	4.5	0.04	0.13	0.06	0.07
japon	7.74	0.85	-0.20	-0.27	-0.11	-0.01	0.9	2.1	0.4	0.0	0.05	0.09	0.01	0.00
---CONTRIBUTION CUMULEE =							7.7	7.7	11.1	46.9	+-----			
GENRE														
1	10.23	0.40	-0.51	0.21	0.20	-0.10	7.6	1.7	2.2	0.6	0.66	0.11	0.11	0.03
2	1.93	6.39	0.46	-2.02	-0.87	0.47	1.1	29.2	7.4	2.3	0.03	0.64	0.12	0.03
3	2.12	5.74	2.06	0.82	-0.19	0.07	25.3	5.3	0.4	0.1	0.74	0.12	0.01	0.00
---CONTRIBUTION CUMULEE =							34.0	36.2	9.9	2.9	+-----			
USAGE														
1	9.73	0.47	-0.53	0.27	0.21	-0.14	7.8	2.7	2.1	1.1	0.61	0.16	0.09	0.04
2	2.81	4.09	0.93	-1.01	-1.05	-0.12	6.9	10.7	15.5	0.2	0.21	0.25	0.27	0.00
3	1.00	13.31	2.37	1.46	0.58	0.57	15.8	7.8	1.7	1.7	0.42	0.16	0.03	0.02
4	0.75	18.08	0.28	-1.69	0.45	1.55	0.2	7.9	0.8	9.5	0.00	0.16	0.01	0.13
---CONTRIBUTION CUMULEE =							30.6	29.1	20.0	12.5	+-----			
COEFB														
0	7.67	0.86	0.36	-0.20	0.13	-0.10	2.8	1.1	0.7	0.4	0.15	0.05	0.02	0.01
1	2.12	5.74	-0.25	0.35	-0.27	-0.59	0.4	1.0	0.8	3.8	0.01	0.02	0.01	0.06
2	4.49	2.18	-0.49	0.17	-0.10	0.44	3.0	0.5	0.2	4.7	0.11	0.01	0.00	0.09
---CONTRIBUTION CUMULEE =							6.2	2.6	1.6	8.9	+-----			
PUISS														
1	4.93	1.90	-0.70	0.19	-0.12	0.76	6.8	0.6	0.4	14.9	0.02	0.01	0.30	0.00
2	4.80	1.97	-0.22	-0.71	0.20	-0.69	0.7	8.9	1.0	12.1	0.03	0.25	0.02	0.24
3	4.55	2.14	1.00	0.54	-0.08	-0.09	12.7	4.9	0.2	0.2	0.46	0.14	0.00	0.00
---CONTRIBUTION CUMULEE =							20.2	14.5	1.5	27.3	+-----			
DURC														
a	11.04	0.29	-0.09	0.11	-0.36	-0.04	0.3	0.5	7.2	0.1	0.03	0.04	0.45	0.01
t	3.24	3.40	0.31	-0.37	1.23	0.14	0.9	1.6	24.6	0.4	0.03	0.04	0.45	0.01
---CONTRIBUTION CUMULEE =							1.2	2.1	31.9	0.5	+-----			

Tableau 2-4 : récapitulatif des modalités bien représentées et à bonnes contributions

	Axe 1	Axe 2	Axe 3	Axe 4
Modalités bien représentées (bons cosinus carrés)	GENRE 1, 3 USAGE 1, 3 PUISS 3	SIN 0,1 GENRE 2 USAGE 2, 3, 4 PUISS 2	SIN 0, 1 USAGE 2 PUISS 1 DURC t, a	MARQUE all, autre PUISS 2
Modalités bien représentées et à contribution significative	GENRE 3, 1 USAGE 3, 1 PUISS 3	GENRE 2 USAGE 2, 3, 4 PUISS 2	USAGE 2 DURC t, a	MARQUE all, autre PUISS 2
Coordonnées positives	GENRE 3 USAGE 3 PUISS 3	USAGE 3	DURC t	MARQUE autre
Coordonnées négatives	GENRE 1 USAGE 1	USAGE 2, 4 GENRE 2 PUISS 2	DURC a USAGE 2	MARQUE all PUISS 2

Tableau 2-5 : Valeurs-tests des modalités significatives (au seuil 5%) de Variables illustratives

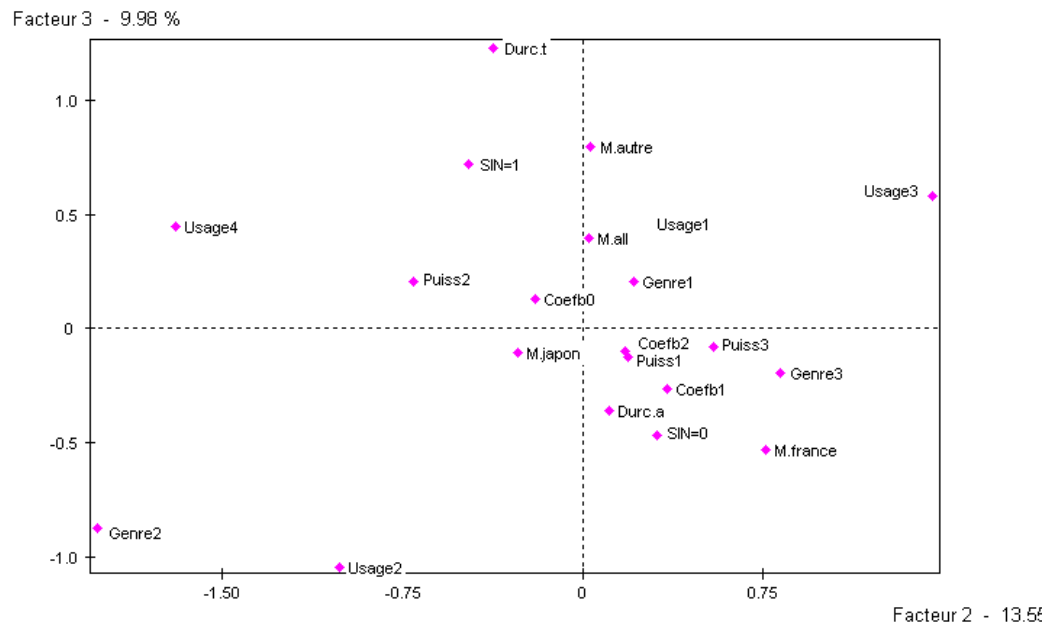
	Modalités	Valeurs-tests
Axe1	VOL 0	4,49
	BRAQ 0	3,38
	VOL 1	-4,49
	BRAQ 1	-3,38
	INC 1	-3,37
Axe2	INC 0	-2,29
	INC 1	2,29
Axe3	VOL 0	3,11
	INC 0	3,19
	VOL 1	-3,11
	INC 1	-3,19
	BRAQ 1	-2,10
Axe4	INC 0	-3,13
	VOL 0	-2,78
	INC 1	3,13
	VOL 1	2,78

Remarque 2-1

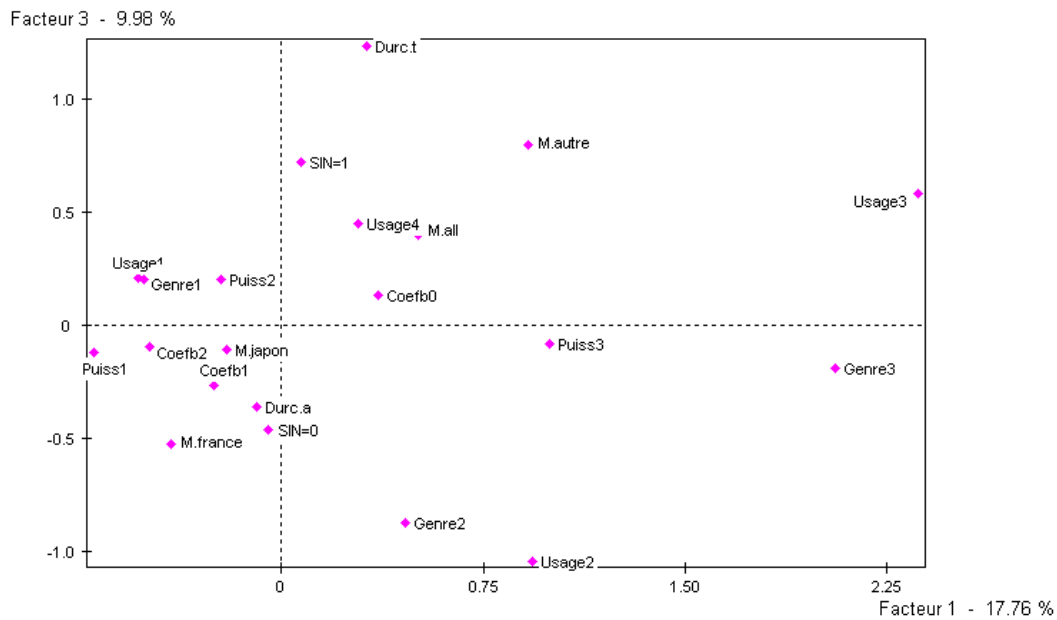
On constate que seuls les axes 2 et 3 nous permettent de dégager certaines caractéristiques de clients à risque. Cependant nous pourrions exploiter les autres axes de façon implicite.

Pour mieux visualiser les interprétations données à nos axes factoriels et les associations ou répulsions pouvant exister entre les différentes modalités, les cartes de modalités dans différents plans factoriels ont été données par les graphiques 2-1 ; 2-2 ; 2-3 ; 2-4 ; et 2-5

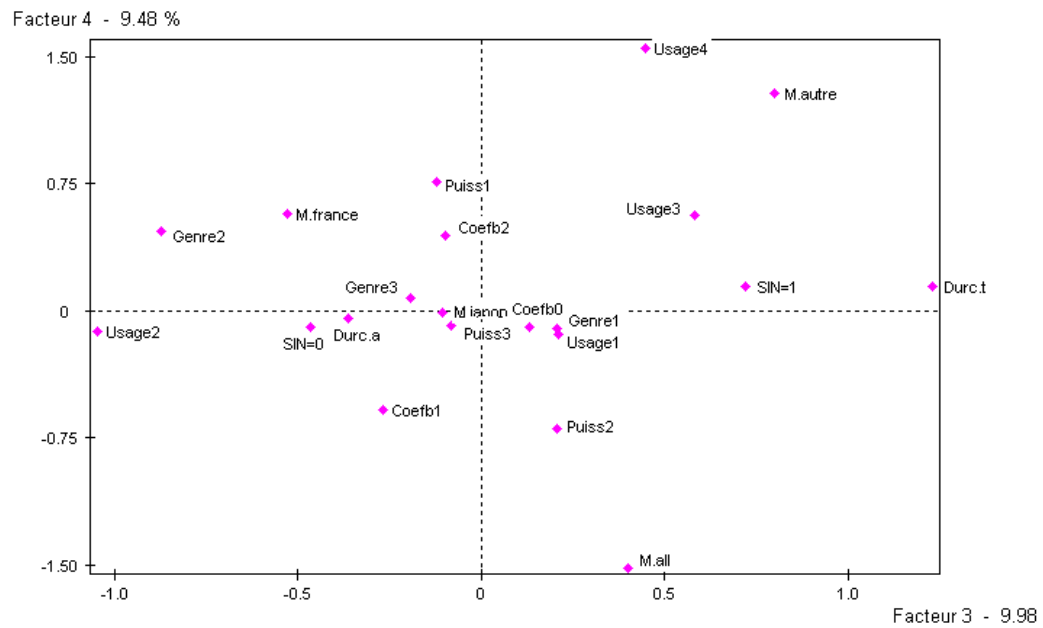
Graphique 2-1 : carte des modalités (axes 2 et 3)



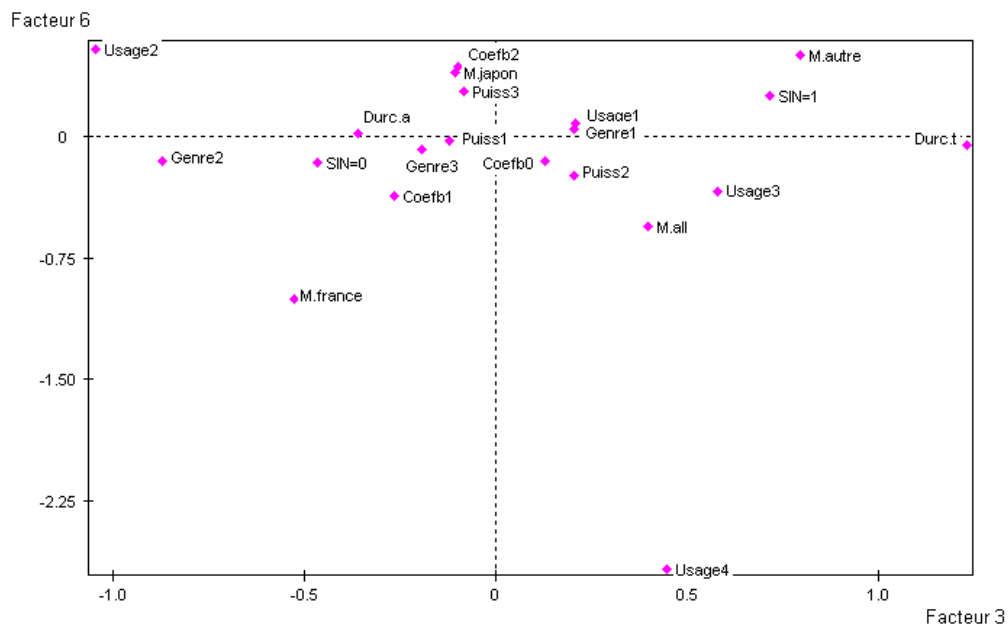
Graphique 2-2 : carte des modalités (axes 1 et 3)

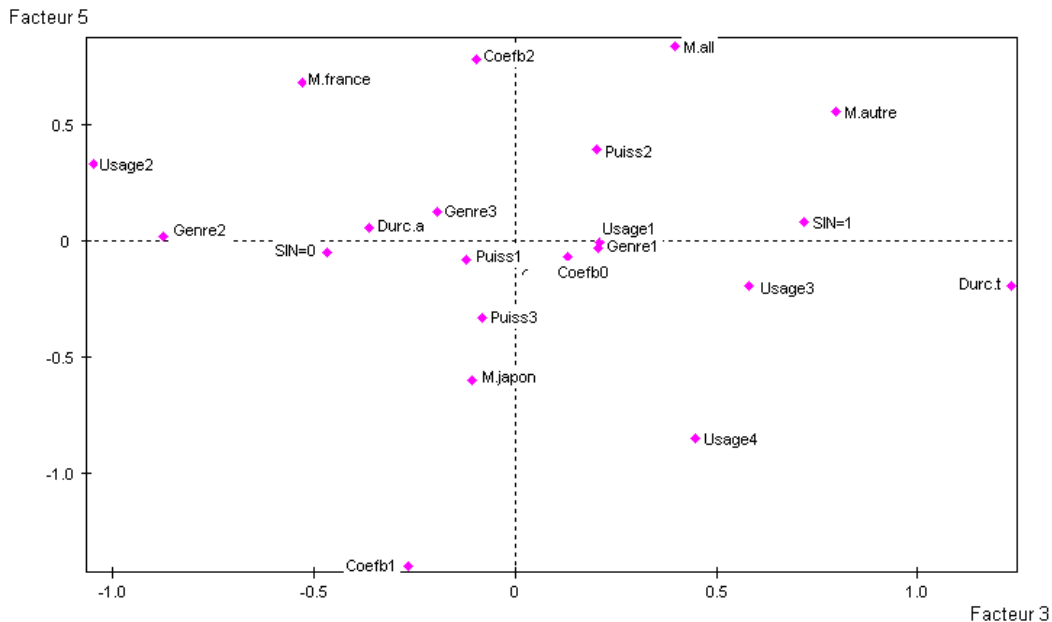


Graphique 2-3 : carte des modalités (axes 3 et 4)



Graphique 2-4 : carte des modalités (axes 3 et 6)



Graphique 2-5 : carte des modalités (axes 3 et 5)**b) Interprétation des graphiques**

Rappelons ici que la proximité entre deux modalités s'interprète en terme d'association entre les modalités.

Le **graphique 2-1** représente le plan factoriel expliquant le mieux la sinistralité. On peut y extraire plusieurs informations, notamment :

- ❖ Les contrats temporaires sont à risque ;
- ❖ Les véhicules de puissance 2 sont plus à risque que ceux de puissance 1 ;

De plus, le rapprochement des modalités GENRE 2 et USAGE 2 nous donne une bonne raison de penser que les variables associées sont corrélées.

Le **graphique 2-2** quant à lui nous montre que :

- ❖ Les véhicules de puissance1 sont plus à risque que ceux de puissance 3 ; et une éventuelle corrélation entre les variables GENRE et PUISS

De façon analogue, le **graphique 2-3** montre que :

- ❖ Les véhicules d'origines « autres » sont plus à risque que ceux d'origines allemandes.

Les véhicules d'origines françaises et japonaises correspondent à des modalités bien représentées sur le sixième axe factoriel (cf. annexe). Ainsi le **graphique 2-4** montre que :

- ❖ Les véhicules d'origines japonaises sont plus à risque que ceux d'origines françaises.

Le **graphique 2-5** nous montre que :

- ❖ Les véhicules d'origines allemandes sont plus à risque que ceux d'origines japonaises.

II - Conclusion

Après interprétation des axes et graphiques, on peut tirer les conclusions suivantes :

- Les contrats temporaires sont à risque ;
- Les véhicules de puissance 2 (puissances fiscales comprises entre 11 et 14 chevaux pour essence ou entre 8 et 10 chevaux pour diesel) sont les plus à risque, suivis de ceux de puissance 1 (puissance fiscale inférieure ou égale à 10 chevaux pour essence ou à 7 chevaux pour diesel), ceux de puissance 3 étant les moins à risque ;
- Pour ce qui est du risque lié à la marque du véhicule, nous pouvons dire que les véhicules de pays d'origines « autres » (ni France, ni Allemagne, ni Japon) sont les plus à risque, ensuite viennent les véhicules d'origine allemande, suivis de ceux d'origine japonaise. On constate donc que les véhicules d'origine Française sont les moins à risque ;
- Les véhicules d'usage 2 sont plus à risque que ceux d'usage 3 ; ce qui pourrait traduire une responsabilité civile plus élevée de la part du conducteur lorsqu'il s'agit d'un transport à titre onéreux de marchandises appartenant à un tiers.

L'information générale que nous retenons des variables supplémentaires est le fait que la « haute sinistralité » ne soit pas essentiellement survenue à l'issue d'un vol, ni d'un incendie, ni d'un braquage. Cependant nous avons déterminé les potentiels souscripteurs à cette garantie ; ce qui représente une étude préliminaire à une analyse des risques essentiellement liés au vol, braquage ou incendie.

Le présent chapitre nous a permis de déceler les modalités se rapprochant le plus de la caractéristique « **haute sinistralité** ». Cependant la question que l'on pourrait se poser est celle de savoir si le rapprochement entre deux modalités ne serait pas dû à l'influence des autres. Ceci est l'objet du prochain chapitre qui a pour principal objectif de séparer les « effets modalités » en retenant celles qui seront les plus pertinentes et les plus discriminantes pouvant expliquer de façon significative la sinistralité.

CHAPITRE TROIS

MODELISATION

Introduction

Nous entendons par modèle un résumé global des relations entre variables, permettant de comprendre des phénomènes, et d'émettre des prévisions.

Rappelons que dans ce chapitre notre but est de proposer le modèle qui s'ajustera le mieux aux données observées ; ceci dans l'intention de mettre en exergue les facteurs décrivant de façon significative la sinistralité. Le but final étant de dire au moyen de ses caractéristiques si le nouveau client est à risque ou non. Dans toute la suite, nous utilisons pour nos différentes analyses le **Logiciel Statistique R**, version 2.1.0 qui date d'avril 2005.

On adoptera le seuil de 10 % comme risque de première espèce pour nos différents tests.

I - Pourquoi le modèle de régression logistique ?

Notre base de données comporte 229 unités statistiques sur lesquelles 15 variables qualitatives ont été enregistrées. La variable à expliquer est la sinistralité, qualitative binaire prenant les niveaux 1 pour « haute sinistralité » et 0 sinon. On est donc en présence d'un événement obéissant à une loi de Bernoulli. Les deux attributs de la variable sinistralité ne pouvant pas être quantifiés de façon naturelle, une idée intuitive de modélisation serait la régression logistique, qui consiste à modéliser la probabilité pour la sinistralité de prendre l'un de ses attributs suivant le vecteur de co-variables observé. D'où le choix judicieux du modèle de régression logistique.

A - Exigences du modèle

Comme tout modèle de régression, le modèle de régression logistique s'applique après vérification de certains critères assurant sa robustesse. On peut entre autre citer les problèmes de sur-dispersion, de sous-dispersion, de colinéarités entre variables explicatives, de liaison entre co-variables et variable réponse.

a) Sur-dispersion et sous-dispersion

Dans le contexte de la régression logistique, si une modalité est le résultat d'un regroupement de plusieurs grappes d'individus, alors il y'a vraisemblablement problème de sur-dispersion. De manière analogue on définit la sous-dispersion. Ces deux notions nous font penser aux variables BG, DOM et IPT de notre base de données.

b) Liaison entre co-variables et variable réponse

Avant d'introduire une variable dans le modèle, il faut d'abord s'assurer de la liaison significative qui existe entre celle-ci et la variable réponse ; au risque de perdre la robustesse du modèle. Nous utiliserons ici le test d'indépendance du khi-deux pour s'assurer des différentes liaisons entre les co-variables et la variable réponse qui est ici la sinistralité.

c) Colinéarités entre les variables explicatives

La colinéarité (ou corrélation) entre deux ou plusieurs variables indépendantes peut affecter la stabilité de leurs coefficients dans le modèle. Plus forte est la corrélation, plus grande est

l'instabilité des coefficients. Pour prévenir ces problèmes d'instabilité, il est recommandé d'inspecter les corrélations qui puissent exister entre les variables indépendantes.

B - Méthodologie de l'analyse

On procède d'abord aux analyses bivariées qui consistent à tester la significativité d'une liaison éventuelle entre la sinistralité (variable à expliquer) et les différentes co-variables décrivant le risque, ceci à l'aide d'un test d'indépendance du Chi deux. En second lieu on fait une analyse multivariée où on prendra en compte les co-variables retenues aux différentes analyses bivariées. Cette analyse multivariée se fera à l'aide du modèle de régression logistique qui nous permettra de faire une étude collective des facteurs décrivant la sinistralité en modélisant la probabilité d'être « haut sinistré » connaissant les caractéristiques du client. Ceci nous permettra de faire des prévisions de statut (« haut sinistré » ou non) pour le nouveau client de la compagnie

a) Analyses bivariées

Avant d'analyser nos données au moyen d'un modèle de régression logistique multivariées, il est d'usage de procéder à des analyses bivariées qui nous permettront d'appréhender les facteurs de risque potentiellement associés avec l'outcome. Sur la base de ces résultats, on procédera à un tri préalable de ces facteurs selon leur degré d'évidence (p-value) et nos connaissances théoriques, afin de ne pas tous les introduire dans le modèle (risque de multi colinéarité, difficulté d'interprétation des résultats, overfitting, etc.). L'analyse se fera ici à l'aide du test d'indépendance du Chi deux dont les principaux résultats sont présentés dans le tableau ci après :

Tableau 3-1 : Analyses Bivariées des facteurs potentiels associés à la sinistralité

covariables	p-value
BG	1.623744e-02 ***
DOM	1.265465e-02 *
VOL	5.651699e-05 ***
HE	9.790238e-01 NS
INC	4.028560e-04 ***
IPT	9.105032e-01 NS
RD	4.287924e-01 NS
BRAQ	4.522924e-05 ***
MARQUE	4.622353e-02 *
GENRE	1.367512e-01 NS
USAGE	7.616820e-01 NS
COEFB	4.199642e-01 NS
PUISS	3.587180e-02 *
DURC	9.785189e-05 ***

Signification des codes : NS : non significatif ; * : significatif à 5% ; *** : très significatif

Commentaire 3-1

Le **tableau 3-1** ci-dessus met en exergue les facteurs potentiels associés à la sinistralité. Le symbole « * » signifie que la variable est associée à la sinistralité ; ainsi les analyses bivariées retiennent les variables :

BG ; DOM ; VOL ; INC ; BRAQ ; MARQUE ; PUISS et DURC comme candidates à l'analyse multivariées.

b) Analyse multivariées

Dans le paragraphe précédent nous avons déterminé les facteurs décrivant de façon significative la sinistralité. Cette fois nous allons considérer un modèle multivariées afin d'étudier l'effet conjoint de plusieurs co-variables sur la probabilité d'être « haut sinistré » ; en essayant d'estimer la contribution pour chaque facteur à l'explication de celle ci. Nous utiliserons pour cela la procédure de sélection pas à pas (implémentée dans le logiciel R) du modèle de régression logistique.

C - Estimation du modèle

a) Un modèle obtenu pas à pas

On commence par le modèle qui contient toutes les co-variables ;
On élimine à chaque étape la co-variable qui a la plus grande **p-value** (probabilité sous l'hypothèse nulle de rejeter l'hypothèse nulle), jusqu'à ce que les co-variables restantes aient une p-value inférieure à une limite donnée. Ici on la prend égale à 0,1 (test au seuil 10 %)

On obtient le premier résultat :

Call:

```
glm(formula = SIN ~ BG + DOM + VOL + INC + BRAQ + MARQUE + PUISS +
    DURC, family = binomial(link = "logit"), data = donnees)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0137	-0.9068	-0.3809	0.9783	2.0801

Coefficients:

	Estimate	Std. Error	z value	p-value
(Intercept)	-0.3513	0.5414	-0.649	0.516451
BG1	-2.2555	1.2616	-1.788	0.073806 .
DOM1	-0.2878	0.8588	-0.335	0.737504
VOL1	-0.9033	0.6041	-1.495	0.134818
INC1	0.2998	0.5937	0.505	0.613594
BRAQ1	-1.3075	0.5710	-2.290	0.022036 *
MARQUEautre	1.5832	0.6266	2.527	0.011509 *
MARQUEfrance	-0.3248	0.5679	-0.572	0.567361
MARQUEjapon	0.5087	0.4526	1.124	0.260999
PUISS2	0.3307	0.4011	0.825	0.409597
PUISS3	-0.9680	0.4293	-2.255	0.024130 *
DURCt	1.2580	0.3784	3.324	0.000886 ***

On supprime dans un premier temps la variable DOM qui a la plus grande p-value et on reprend le procédé. On obtient alors le modèle comprenant les variables : BG ; VOL ; BRAQ ; MARQUE ; PUISS et DURC

Une analyse de colinéarité entre ces variables est donnée par le tableau ci-dessous :

Tableau 3-2 : Analyse de colinéarité par le test du Chi deux

variables		p-value
VOL	BRAQ	1.835e-13 ***
BRAQ	BG	7.684e-05 ***
BG	VOL	0.003548 ***
DURC	BRAQ	0.05408 °

°: significatif à 10% *** : très significatif

Remarque 3-1

On remarque à l'aide du tableau ci-dessus que les variables décrivant les garanties vol, bris de glace et braquage sont très corrélées deux à deux. Il serait donc redondant de les considérer toutes dans un même modèle, car une corrélation entre deux variables signifie que les deux variables apportent presque la même information. Ainsi pour chaque paire, le choix de la variable à retenir se fera en fonction du degré de liaison avec la sinistralité. De ce fait nous retenons dans un premier temps la variable BRAQ comme représentative des garanties souscrites. Cependant on constate que celle-ci est corrélée à la variable indiquant le type de contrat souscrit (DURC). Dans la suite nous retenons la variable DURC pour sa pertinence.

On retient donc au final le **modèle 1** :

Call:

```
glm(formula = SIN ~ PUISS + DURC, family = binomial(link = "logit"),
    data = donnees)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6142	-0.9032	-0.6947	1.1578	1.7550

Coefficients:

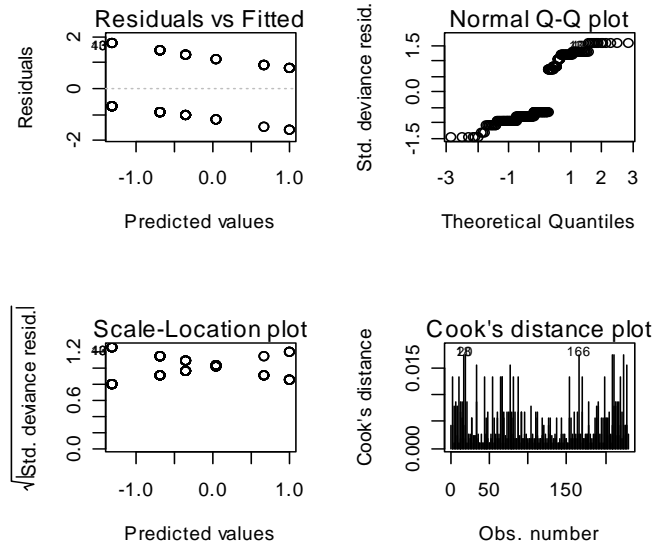
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.6858	0.2458	-2.790	0.00526 **
PUISS2	0.3266	0.3366	0.970	0.33191
PUISS3	-0.6128	0.3632	-1.687	0.09156 .
DURCt	1.3449	0.3393	3.964	7.36e-05 ***

Null deviance: 306.90 on 228 degrees of freedom

Residual deviance: 283.58 on 225 degrees of freedom

AIC: 291.58

Le modèle 1 ci-dessus est admissible car les variables sont toutes significatives au seuil indiqué (10%). Cependant il est important pour nous d'inspecter son graphique de diagnostic.

Graphique 3-1 : Quelques graphes de diagnostic du modèle 1**Remarque 3-2**

Le graphe des coefficients de Cook [5] montre que certaines observations ont des coefficients de Cook très importants, notamment les observations 18, 20 et 166 qui pourraient être des outliers (valeurs aberrantes). Il convient d'ailleurs de rappeler que le graphe des distances de Cook (**Cook's distance plot**) mesure pour un individu l'écart entre la valeur observée et celle prédite par le modèle. Une distance trop grande signifie donc que l'ajustement n'est pas correct en ce point.

Essayons maintenant de supprimer les enregistrements 18, 20 et 166 de notre base de données. Pour le faire nous utiliserons une procédure R d'extraction automatique de données.

Après cette suppression, le graphique des distances de Cook montre à nouveau trois outliers : les enregistrements 17, 18, et 163. Ainsi nous supprimons au total six enregistrements de la base de données initiale. Notre base de données comporte désormais 223 enregistrements. Le modèle ajusté à cette nouvelle base est donné ci-dessous ; appelons le « **modèle 2** ».

```
glm(formula = SIN ~ PUISS + DURC, family = binomial(link = "logit"),
    data = donnees)
```

Coefficients:

	Estimate	Std.Error	z value	Pr(> z)
(Intercept)	-0.6707	0.2497	-2.686	0.00722 **
PUISS2	0.3383	0.3432	0.986	0.32433
PUISS3	-0.6362	0.3671	-1.733	0.08313 .
DURCt	1.3703	0.3472	3.947	7.93e-05 ***

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 299.16 on 222 degrees of freedom
 Residual deviance: 275.64 on 219 degrees of freedom
 AIC: 283.64

Remarque 3-3

On constate que le modèle obtenu après suppression de quelques valeurs aberrantes (**Modèle 2**) s'ajuste nettement mieux aux données, comparé au **modèle 1**. On peut le voir en comparant les critères d'information d'Akaike [2](AIC) associés aux deux modèles ; l'ajustement étant d'autant plus bon que son AIC est faible. Cet écart considérable entre les deux critères d'information (**291,58** et **283,64**) nous permet de voir l'effet néfaste que pourraient apporter ces valeurs aberrantes dans notre modélisation. On s'intéressera donc pour la suite au **modèle 2**.

b) Interprétation des coefficients du modèle

Les coefficients du modèle s'interprètent bien :

- Les contrats temporaires sont corrélés positivement avec la sinistralité ; c'est à dire que les contrats temporaires sont à risque. De plus le symbole '***' nous montre que le degré de significativité du test de nullité du coefficient associé à cette variable est assez élevé ; ce qui signifierait que la quasi totalité des contrats temporaires sont à haut risque.
- Les véhicules de puissances fiscales supérieures à 15 chevaux (essence) et 11 chevaux (diesel) sont corrélés négativement avec la sinistralité. Ceci signifie que les véhicules de pareilles puissances fiscales sont les moins à risque, car le signe '-' signifie que cette caractéristique diminue la probabilité d'être « haut sinistré ». Nous reviendrons sur l'expression explicite de cette probabilité.

Les analyses effectuées jusqu'ici nous ont permis d'adopter le **modèle 2** comme modèle final.. Il s'interprète comme suit :

La probabilité d'être « haut sinistré » connaissant la puissance fiscale du véhicule et le type de contrat souscrit (temporaire ou annuel) peut être estimée par :

$$\Pr(SIN = 1 / PUISS, DURC) = \frac{e^{\theta}}{1 + e^{\theta}} ;$$

Avec : $\theta = -0.6707 - (0.6362)1_{PUISS} + (1.3703)1_{DURC}$

$1_{PUISS} = 1$ si le véhicule a une puissance fiscale supérieure à 15 chevaux (pour essence)
ou 11 chevaux (pour diesel) ;
0 sinon

$1_{DURC} = 1$ si le contrat est temporaire (pas annuel) ;
0 sinon

D - Validation du modèle

Maintenant que nous avons retenu un modèle, il reste à le valider ; c'est à dire à mesurer son ajustement à notre base de données et sa capacité de prédiction pour les nouveaux clients.

Un indicateur d'ajustement d'un modèle de régression logistique est basé sur sa déviance résiduelle : pour un modèle bien ajusté, la déviance résiduelle divisé par son degré de liberté doit être approximativement égal à 1 [7].

Pour le modèle retenu la déviance résiduelle vaut 275.64 correspondant à 219 degrés de liberté ; ce qui donne un rapport égal à 1,25

Il existe en pratique plusieurs méthodes de validation de modèle, parmi lesquelles la validation croisée qui est vivement recommandée. Cette méthode de validation consiste à faire une

répartition aléatoire des observations en deux parties et à appliquer à une partie des observations le modèle construit sur l'autre partie des observations.

Dans le cas présent, compte tenu du volume réduit de notre base de données, nous ne pourrions pas utiliser la méthode de validation croisée ; mais plutôt la méthode consistant à tester le modèle sur les données qui ont servi pour son estimation. Nous y reviendrons dans le prochain paragraphe.

a) Evaluation du pouvoir prédictif du modèle

Nous avons jusque là vu que la régression logistique permet d'estimer la probabilité d'être « haut sinistré » ($SIN=1$) quand on connaît la puissance fiscale du véhicule et le type de contrat souscrit. Sur la base de ces probabilités, on peut définir une règle de classification de la manière suivante :

- ❖ Si la probabilité est supérieure à un seuil S_0 fixé, on classe le client comme « haut sinistré » ($SIN=1$)
- ❖ Si au contraire la probabilité est inférieure ou égale à S_0 , le client n'est pas classé comme « haut sinistré » ($SIN=0$)

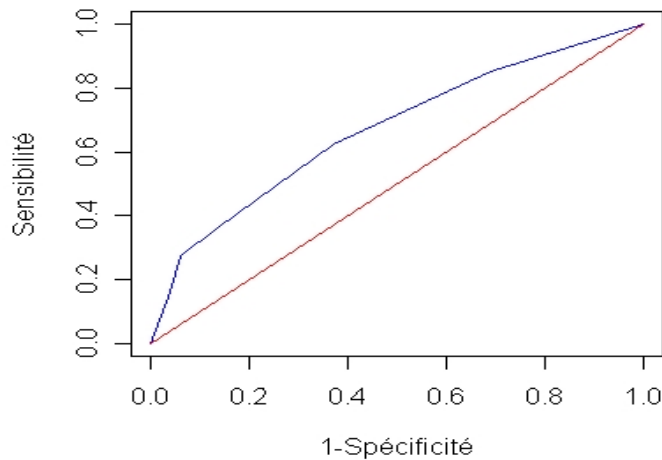
Bien que le seuil $S_0 = 0.5$ paraisse a priori une valeur raisonnable, il n'est pas du tout évident que ce soit exact. Pour chaque valeur de S_0 on peut calculer la sensibilité et la spécificité du modèle. La sensibilité est définie comme la probabilité de classer l'individu dans la catégorie $SIN = 1$ étant donné qu'il est effectivement observé dans celle-ci :

$$\text{Sensibilité} = P_r(\text{« haut sinistré »} \mid SIN = 1)$$

La spécificité quant à elle est la probabilité de classer l'individu dans la catégorie $SIN = 0$ étant donné qu'il est effectivement observé dans celle-ci :

$$\text{Spécificité} = P_r(\text{« non haut sinistré »} \mid SIN = 0)$$

La qualité de la méthode de classification est généralement mesurée par ces deux indicateurs (sensibilité et spécificité) au moyen de la courbe ROC (Receiver Operating Characteristic curve) qui est la courbe représentative de la sensibilité en fonction de (1-spécificité). Ainsi, l'aire au dessous de la courbe ROC nous permet de mesurer globalement la capacité du modèle à affecter correctement les sujets à leurs classes respectives. Le **graphique 3-2** ci-dessous donne la courbe ROC correspondante au modèle retenu (modèle 2).

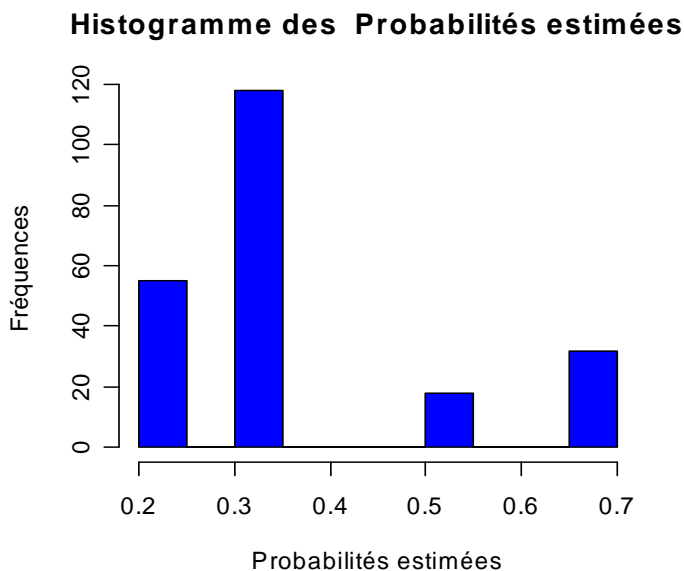
Graphique 3-2 : Courbe ROC

Ce graphique représente une aire au dessous de la courbe ROC $C_0 = 0.672$. Rappelons que ce coefficient C_0 représentant l'aire au dessous de la courbe ROC pour notre modèle n'est pas loin du critère $0.7 \leq C < 0.8$ correspondant à une discrimination acceptable [8].

L'aire au dessous de la courbe ROC nous montre que la discrimination n'est pas très bonne ; ceci pouvant s'expliquer par l'existence d'un groupe important d'individus ayant des statuts différents, (« hauts sinistrés » ou non) mais de profils semblables. Pour remédier à une pareille situation, il est recommandé d'utiliser une base de données de volume important (de l'ordre des milliers d'enregistrements)

b) Choix de la probabilité seuil (S_0)

Avant de choisir le seuil de probabilité S_0 à adopter pour les nouveaux clients de la compagnie, le graphique 3-3 ci-dessous nous donne l'histogramme des différentes probabilités estimées par le modèle.

Graphique 3-3 :

Cet histogramme (graphique3-3) montre qu'en prenant le seuil $S_0 = 0.5$, une proportion très faible de sinistrés de notre base de données se retrouverait dans la classe des « hauts sinistrés ». Le modèle prédira exactement 50 « hauts sinistrés » au lieu de 88 que comporte la base de données. D'où la nécessité du choix d'un seuil S_0 un peu plus permissif.

Le choix du seuil de probabilité S_0 ne se fait pas de façon arbitraire. Pour cela, construisons le test d'hypothèses statistiques :

(H₀) : Le client n'est pas « haut sinistré » contre (H₁) : Le client est « haut sinistré »

Fixons comme risque de première espèce du test $\alpha \in]0, 1[$; c'est à dire la probabilité pour le modèle de prédire à tort les « non hauts sinistrés » vaut α

Dans cette étude nous modélisons la probabilité d'être haut sinistré. De ce fait, nous pouvons prendre comme statistique de test les probabilités estimées par le modèle. Soit Z cette statistique de test ; et z une réalisation de la variable aléatoire Z .

Notre test d'hypothèses (H₀) contre (H₁) se formule comme suit :

- Si $z \leq S_\alpha$, on accepte (H₀)
- Si $z > S_\alpha$, on rejette (H₀)

Avec S_α choisi de sorte que la probabilité pour le modèle de mal prédire le client « non haut sinistré » soit égale à α ;

C'est à dire $\Pr_{H_0}(Z > S_\alpha) = \alpha$

D'où $F_Z^0(S_\alpha) = 1 - \alpha$

F_Z^0 étant la fonction de répartition de la loi de Z sous l'hypothèse (H₀)

S_α est donc le $(1-\alpha)$ -quantile de la loi de Z dans la sous-population des « non hauts sinistrés »

Ne connaissant pas à priori la loi de Z nous devons estimer S_α de manière empirique. Pour ce faire, plaçons nous dans la sous-population des « non hauts sinistrés » et examinons les différentes probabilités estimées par le modèle. Ainsi nous obtenons le tableau ci-dessous :

Tableau 3-3 : Fréquences des Probabilités estimées par le modèle sous (H₀)

Probabilités estimées sous (H ₀)	Fréquences en %	Fréquences cumulées en %
0.213	31.1	31.1
0.338	55.6	86.7
0.515	7.4	94.1
0.668	5.9	100

De ce tableau nous pouvons lire l'approximation : $F_Z^0(0.338) \approx 0.867$; ce qui nous permet d'avoir les estimations : $1 - \hat{\alpha} = 0.867$ et $\hat{S}_\alpha = 0.338$

- Puissance du test ($\pi(\alpha)$)

Pour un risque de première espèce α fixé, on définit la puissance π_α du test associé par :

$$\pi_\alpha = 1 - \beta ;$$

β étant la probabilité pour le modèle de faire une mauvaise prédiction pour le client « haut sinistré » ; c'est à dire $\beta = \Pr_{H_1}(Z \leq S_\alpha)$. β est encore appelé risque de deuxième espèce.

De ce fait nous pouvons définir la puissance π_α du test comme étant la probabilité pour le modèle de bien prédire les « hauts sinistrés ».

Le calcul de cette puissance pour la valeur de α obtenue ci-dessus ($\alpha = 0.133$) nous emmène à considérer la sous-population des « hauts sinistrés » de notre base de données. Le tableau 3-4 donne un résumé des probabilités estimées dans cette sous-population.

Tableau 3-4 : Fréquences des Probabilités estimées par le modèle sous (H_1)

Probabilités estimées sous (H_1)	Fréquences en %	Fréquences cumulées en %
0.213	14.8	14.8
0.338	48.9	63.7
0.515	9	72.7
0.668	27.3	100

Remarque 3-4

Ce tableau nous permet d'obtenir l'approximation : $\hat{\beta} = 63.7\%$; c'est à dire $\hat{\pi}_\alpha = 36.3\%$

Rappelons que cette puissance de test n'est pas vraiment satisfaisante ; ceci pouvant s'expliquer par le fait qu'une proportion considérable de « hauts sinistrés » (48.9 %) se retrouve au seuil de probabilité $S_{13,3\%} = 0.338$. D'où la nécessité de création d'une zone d'indifférence correspondante à la probabilité seuil.

- Règle de décision

Tout ce qui a été fait jusque là est dans le but de pouvoir dire pour un nouveau client de la compagnie, s'il est a priori à « haut risque » ou pas. Ainsi nous pouvons définir la règle de classification de la manière suivante :

Si $z \leq 0.338$; on a de bonnes raisons de penser que le client ne soit pas à « haut risque » ;

Si $z > 0.338$; on a de bonnes raisons de penser que le client soit à « haut risque » ;

z étant une réalisation de la variable aléatoire Z représentant les probabilités estimées par le modèle.

c) Erreur de prédiction

La règle de décision élaborée ci-dessus permet de détecter 36.3 % de « hauts sinistrés » et 86.7 % de « non hauts sinistrés ».

L'erreur de prédiction associée à cette règle est la probabilité de faire une mauvaise affectation. Soit ε cette erreur de prédiction ; on obtient :

$$\varepsilon = \Pr(\text{« haut sinistré », la règle décide le contraire}) + \Pr(\text{« Non haut sinistré », la règle décide le contraire})$$

$$\begin{aligned}
&= \Pr(SIN = 1, Z \leq 0.338) + \Pr(SIN = 0, Z > 0.338) \\
&= \Pr(Z \leq 0.338 / SIN = 1) \Pr(SIN = 1) + \Pr(Z > 0.338 / SIN = 0) \Pr(SIN = 0) \\
&= \Pr_{H_1}(Z \leq 0.338) \Pr(SIN = 1) + \Pr_{H_0}(Z > 0.338) \Pr(SIN = 0)
\end{aligned}$$

On peut faire les estimations (à partir de notre base de données) :

$$\Pr(SIN = 1) = \frac{88}{223} \approx 0.394 \quad ; \quad \Pr(SIN = 0) = \frac{135}{223} \approx 0.605 \quad ; \quad \Pr_{H_1}(Z \leq 0.338) = \beta \approx 63.7\% \quad ;$$

$$\Pr_{H_0}(Z > 0.338) = \alpha \approx 13.3\%$$

Ainsi nous obtenons : $\varepsilon = 33\%$

Remarque 3-5

Rappelons qu'une démarche plus raisonnable de calcul d'erreur de prédiction consiste à utiliser les données n'ayant pas servi à l'estimation du modèle [11]. Mais compte tenu du volume réduit de la base de données, nous l'avons calculé à partir des données d'apprentissage.

Le fait d'avoir une proportion considérable (48.9 %) de clients à haut risque correspondant au seuil de probabilité S_0 et déclarés « non hauts sinistrés » par la règle de décision nous amène à considérer une classe « tampon » ; dite d'indécision. Cette classe correspond aux clients ayant des contrats annuels et dont les véhicules sont de puissances fiscales inférieures à 15 chevaux (Essence) ou 11 chevaux (Diesel). Rappelons que **l'adoption de pareils clients dans la classe définie par le modèle est une opération dangereuse ; au risque d'une éventuelle tombée en faillite de la compagnie**. Afin de trouver un critère objectif d'affectation de clients appartenant à cette classe, nous devons mettre les clients qui la constituent en observation pendant un certain temps, et faire une analyse de données censurées en utilisant par exemple le modèle de régression de Cox [9] et comparer les estimateurs de Kaplan-Meier [9]. Avant que cela ne soit fait, nous proposons une règle de décision finale.

d) règle de décision finale

Si $z < 0.338$; on a de bonnes raisons de penser que le client ne soit pas à « haut risque » ;

Si $z > 0.338$; on a de bonnes raisons de penser que le client soit à « haut risque » ;

Si $z = 0.338$; on est dans la classe d'indécision

Ceci nous permet d'obtenir à l'aide du logiciel R un programme de classification automatique prenant en entrée la puissance fiscale du véhicule et le type de contrat.

e) Programme R de Classification Automatique

```

Score =function(PUISS,DURC){
puissance3 = ifelse(PUISS==3,1,0)
temporaire = ifelse(DURC=="t",1,0)
theta = -0.6707-0.6362*puissance3+1.3703*temporaire
score = exp(theta)/(1+exp(theta))
if (score < 0.338) { paste(« le client n'est pas à haut risque ») }
else { if (score > 0.338) { paste(« le client est à haut risque ») }
      else { paste(« le client est dans la zone d'indécision ») }
    }
}

```

II - Conclusion

Ce chapitre nous a permis dans un premier temps de déterminer un groupe de variables pouvant expliquer de manière significative (au seuil 10 %) la sinistralité. L'analyse simultanée de ce groupe de variables par le modèle logistique ressort les plus pertinentes et plus discriminantes pouvant représenter valablement le groupe : la durée du contrat et la puissance fiscale du véhicule. Après examen des résultats de l'analyse, on peut noter que :

- Les contrats **temporaires** représentent un « **gros risque** » pour la compagnie ;
- Les véhicules de **puissance fiscale inférieure à 15 chevaux (pour essence) ou 11 chevaux (pour diesel)** sont les plus responsables de « gros sinistres » ;
- Le pays d'origine du véhicule est associé à la sinistralité; les véhicules d'origines « **autres** » étant les plus exposés au risque.

D'autre part, nous avons construit un test statistique pour la détermination du seuil de probabilité à partir duquel la discrimination sera faite. Cependant rappelons que ce seuil de probabilité correspond à une classe de clients dite « d'indécision », c'est à dire pour laquelle l'affectation est ambiguë. Pour remédier à cela nous proposons une mise en observation de pareils clients pendant un certain temps, pour une analyse de données censurées qui conduira à des critères plus objectifs d'affectation des individus dans l'une des deux classes. Avant que cela ne soit fait, un programme R de classification automatique a été proposé.

Conclusion Générale

L'objectif de notre étude était de déterminer le profil des clients à « haut risque » du portefeuille automobile de la compagnie d'assurance **CHANAS ASSURANCES S.A**, antenne de Yaoundé. Cette notion de « haut risque » a été définie au début de l'étude.

Pour mener à bien notre étude, nous avons d'abord exploré la base de données qui nous a été confiée par la compagnie, puis nous avons construit un modèle s'ajustant aux données observées et permettant d'affecter sans grand risque de se tromper les nouveaux clients de la compagnie dans l'une des classes.

Les analyses bivariées par le test d'indépendance du Chi deux montrent que : Le type de contrat souscrit (temporaire ou annuel), la puissance fiscale du véhicule et le pays d'origine du véhicule sont les facteurs pertinents significativement liés à la sinistralité.

De l'analyse des correspondances multiples (ACM) il ressort dans un premier temps que les **véhicules de puissances fiscales comprises entre 11 et 14 chevaux (Essence) ou entre 8 et 10 chevaux (Diesel) sont les plus à risque** ; suivis de ceux de puissance fiscale inférieure ou égale à 10 chevaux (Essence) ou à 7 chevaux (Diesel). De même, l'ACM montre que **les véhicules de pays d'origines « autres »** (ni France, ni Japon, ni Allemagne) sont les plus responsables de « gros sinistres ». Ensuite viennent les véhicules d'origine Allemande, puis ceux d'origine Japonaise. Les véhicules d'origine Française étant les moins à risque.

L'analyse multivariée par le modèle de régression logistique nous a permis d'une part de séparer les colinéarités existantes du fait de l'influence de certains facteurs sur les autres ; ceci en proposant deux prédicteurs pertinents qui pourraient représenter valablement tous les autres. Il s'agit en fait du type de contrat souscrit (temporaire ou annuel) et de la puissance fiscale du véhicule. On note ici que **les contrats temporaires augmentent considérablement la sinistralité**. D'autre part, par le biais du modèle logistique, nous avons construit un programme **R** permettant à l'assureur d'affecter le nouveau client dans l'une des classes (« hauts sinistrés » ou « non hauts sinistrés ») sans grand risque de se tromper. Cependant une zone d'indécision a été construite

Toutefois, nos résultats ne pourront pas servir de manière absolue comme outil de référence pour le décideur, compte tenu de la taille réduite de notre échantillon d'étude ; ce qui nous a d'ailleurs mis en désaccord avec l'application de la méthode de validation croisée du modèle.

Perspectives et Recommandations

La phase la plus difficile de notre stage a été la collecte des données qui s'est faite essentiellement sur support physique. A cet effet il serait souhaitable pour la compagnie d'améliorer son système d'information en mettant sur pied un système complet d'archivage des données numériques. Pour cela, l'adoption d'un système de gestion de base de données (SGBD) un peu plus sophistiqué serait préférable, car il permettrait d'avoir une base de données beaucoup plus exploitable.

Une des restrictions de cette étude est due au fait que certaines variables permettant de mieux renseigner la sinistralité n'ont pas été prises en compte [6]. Nous pouvons entre autres citer l'âge du véhicule, celui du conducteur habituel du véhicule, son statut matrimonial, sa nationalité et pour le service des sinistres, l'âge du conducteur au moment du sinistre. Le service production de la compagnie devrait donc mettre à profit les fiches de souscription de contrat ; en veillant à ce que toutes ces informations soient enregistrées pendant la souscription.

Dans cette étude nous avons comme population statistique quelques sinistrés du portefeuille automobile de la compagnie. L'analyse multivariées par le modèle de régression logistique nous a permis de faire des prévisions pour des clients susceptibles d'être à « haut risque » ou non. La question que l'on pourrait se poser est celle de savoir si pour un nouveau client il est normal de le qualifier de client à « haut risque » ne sachant pas s'il verra la survenance d'un sinistre avant la fin de son contrat ?

Pour rendre l'étude plus intéressante, il serait souhaitable pour la compagnie de faire une étude préliminaire consistant à modéliser la probabilité de survenance d'un sinistre en fonction des caractéristiques du client. Ceci permettra à l'assureur de savoir a priori les chances de survenance de sinistre pour le nouveau client de la compagnie.

Après cette étude préliminaire, la compagnie devrait entreprendre une étude semblable à celle que nous venons de mener, cette fois avec un volume d'information plus consistant (de l'ordre des milliers) ; afin de mieux explorer sa base de données et d'y extraire des informations permettant de faire une tarification conséquente du risque, et surtout d'assurer la stabilité des provisions mathématiques.

Par ailleurs les archives du service production de la compagnie font état d'un nombre assez élevé de contrats résiliés. A cet effet une étude statistique serait nécessaire afin de déceler les principaux facteurs qui pourraient être à l'origine de ces résiliations ; ceci dans le but de remédier à cette situation par l'utilisation des techniques conséquentes de marketing.

En définitive, comme suggestion par rapport aux résultats de notre analyse, la compagnie devrait adopter une politique de gestion tendant à **réduire au maximum les contrats temporaires**. En ce qui concerne les clients à « haut risque », une re-tarification s'avère nécessaire ; celle-ci prenant en compte les différents résultats de l'analyse.

ANNEXES

I- Principaux Programmes R utilisés

```
# -----lecture des données-----#

donnees=read.table("chanas.sa.txt",header=TRUE)

#-----fonction qui transforme les variables en facteur-----#

factor=function(d) {
  for (j in 1:15){
    d[,j]=as.factor(d[,j])
  }
  d
}

#----- Analyses Bivariables par le test d'indépendance du Chi-deux-----#

chideux=function(d){
  p=vector()
  for(j in 2:15){
    p[j-1]=(chisq.test(d[,j],SIN))[[3]]
  }
  p
}

#-----Modèle retenu-----#

donnees=read.table("chanas.sa.txt",header=TRUE)
donnees=factor(donnees)
attach(donnees)
modele=glm(formula=SIN~DURC+PUISS,family=binomial(link="logit"),data=donnees)
summary(modele)
#----- Aire au dessous de la courbe ROC et Courbe ROC -----#

library(Design)
mod=lrn(formula=SIN~DURC+PUISS,x=T,y=T,data=donnees)
mod
Logistic Regression Model

      Obs  Max Deriv Model L.R.  d.f.   P     C   Dxy
      223  5e-12   23.52     3     0  0.672 0.344

library(ROCR)
p=predict(modele)
pre=prediction(p,SIN)
perf1 <- performance(pre, "tpr", "fpr")
plot(perf1,col='blue')
plot(perf1,col='blue',xlab='1-Spécificité',ylab='Sensibilité')
lines(c(0,1),c(0,1),col='red')

#-----Histogramme des probabilités estimées par le modèle-----#
```

```

Proba_est=function(x,y){
proba.est=vector()
for(i in 1:223){
a=-0.6707-0.6362*as.numeric(x[i])+1.3703*as.numeric(y[i])
proba.est[i]=exp(a)/(1+exp(a))
}
proba.est
}
hist(Proba_est(ifelse(PUISS==3,1,0),ifelse(DURC=="t",1,0)),col="blue",main="Histogramme des
Probabilités estimées",xlab="Probabilités estimées",ylab="Fréquences")

```

II- Listings des résultats d'ACM avec le logiciel Spad version 4.01

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRÉS DES MODALITES ACTIVES

AXES 1 A 5

LIBELLE	MODALITES		COORDONNEES					CONTRIBUTIONS					COSINUS CARRÉS				
	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1 . SIN																	
0	8.67	0.65	-0.05	0.31	-0.47	-0.10	-0.05	0.1	3.1	9.4	0.4	0.2	0.00	0.15	0.33	0.01	0.00
1	5.61	1.54	0.07	-0.48	0.72	0.15	0.08	0.1	4.7	14.5	0.6	0.2	0.00	0.15	0.33	0.01	0.00
CONTRIBUTION CUMULEE = 0.1 7.8 23.9 1.1 0.4 +-----																	
10 . MARQUE																	
all	2.31	5.19	0.51	0.03	0.40	-1.52	0.84	1.7	0.0	1.8	28.3	10.5	0.05	0.00	0.03	0.45	0.14
autre	1.62	7.81	0.92	0.03	0.80	1.29	0.56	3.9	0.0	5.2	14.1	3.3	0.11	0.00	0.08	0.21	0.04
france	2.62	4.45	-0.41	0.76	-0.53	0.57	0.68	1.3	5.6	3.7	4.5	7.8	0.04	0.13	0.06	0.07	0.10
japon	7.74	0.85	-0.20	-0.27	-0.11	-0.01	-0.60	0.9	2.1	0.4	0.0	17.9	0.05	0.09	0.01	0.00	0.42
CONTRIBUTION CUMULEE = 7.7 7.7 11.1 46.9 39.5 +-----																	
11 . GENRE																	
1	10.23	0.40	-0.51	0.21	0.20	-0.10	-0.03	7.6	1.7	2.2	0.6	0.1	0.66	0.11	0.11	0.03	0.00
2	1.93	6.39	0.46	-2.02	-0.87	0.47	0.02	1.1	29.2	7.4	2.3	0.0	0.03	0.64	0.12	0.03	0.00
3	2.12	5.74	2.06	0.82	-0.19	0.07	0.12	25.3	5.3	0.4	0.1	0.2	0.74	0.12	0.01	0.00	0.00
CONTRIBUTION CUMULEE = 34.0 36.2 9.9 2.9 0.3 +-----																	
12 . USAGE																	
1	9.73	0.47	-0.53	0.27	0.21	-0.14	-0.01	7.8	2.7	2.1	1.1	0.0	0.61	0.16	0.09	0.04	0.00
2	2.81	4.09	0.93	-1.01	-1.05	-0.12	0.33	6.9	10.7	15.5	0.2	2.0	0.21	0.25	0.27	0.00	0.03
3	1.00	13.31	2.37	1.46	0.58	0.57	-0.19	15.8	7.8	1.7	1.7	0.2	0.42	0.16	0.03	0.02	0.00
4	0.75	18.08	0.28	-1.69	0.45	1.55	-0.85	0.2	7.9	0.8	9.5	3.5	0.00	0.16	0.01	0.13	0.04
CONTRIBUTION CUMULEE = 30.6 29.1 20.0 12.5 5.7 +-----																	
13 . COEPB																	
0	7.67	0.86	0.36	-0.20	0.13	-0.10	-0.07	2.8	1.1	0.7	0.4	0.2	0.15	0.05	0.02	0.01	0.01
1	2.12	5.74	-0.25	0.35	-0.27	-0.59	-1.40	0.4	1.0	0.8	3.8	26.9	0.01	0.02	0.01	0.06	0.34
2	4.49	2.18	-0.49	0.17	-0.10	0.44	0.78	3.0	0.5	0.2	4.7	17.7	0.11	0.01	0.00	0.09	0.28
CONTRIBUTION CUMULEE = 6.2 2.6 1.6 8.9 44.9 +-----																	
14 . PUISS																	
1	4.93	1.90	-0.70	0.19	-0.12	0.76	-0.08	6.8	0.6	0.4	14.9	0.2	0.26	0.02	0.01	0.30	0.00
2	4.80	1.97	-0.22	-0.71	0.20	-0.69	0.39	0.7	8.9	1.0	12.1	4.8	0.03	0.25	0.02	0.24	0.08
3	4.55	2.14	1.00	0.54	-0.08	-0.09	-0.33	12.7	4.9	0.2	0.2	3.2	0.46	0.14	0.00	0.00	0.05
CONTRIBUTION CUMULEE = 20.2 14.5 1.5 27.3 8.3 +-----																	
15 . DURC																	
a	11.04	0.29	-0.09	0.11	-0.36	-0.04	0.06	0.3	0.5	7.2	0.1	0.2	0.03	0.04	0.45	0.01	0.01
t	3.24	3.40	0.31	-0.37	1.23	0.14	-0.19	0.9	1.6	24.6	0.4	0.8	0.03	0.04	0.45	0.01	0.01
CONTRIBUTION CUMULEE = 1.2 2.1 31.9 0.5 1.0 +-----																	

AXES 6 A 10

LIBELLE	MODALITES		COORDONNEES							CONTRIBUTIONS					COSINUS CARRÉS				
	P.REL	DISTO	6	7	8	9	10	6	7	8	9	10	6	7	8	9	10		
1 . SIN																			
0	8.67	0.65	-0.16	0.30	0.17	0.21	-0.13	1.6	5.7	2.0	3.4	1.4	0.04	0.13	0.05	0.07	0.02		
1	5.61	1.54	0.25	-0.46	-0.27	-0.32	0.19	2.5	8.8	3.1	5.2	2.1	0.04	0.13	0.05	0.07	0.02		
CONTRIBUTION CUMULEE = 4.2 14.6 5.1 8.6 3.4 +-----																			
10 . MARQUE																			
all	2.31	5.19	-0.56	0.15	0.26	0.03	-0.65	5.1	0.4	1.2	0.0	9.6	0.06	0.00	0.01	0.00	0.08		
autre	1.62	7.81	0.51	0.09	1.35	-1.20	-0.36	3.0	0.1	23.2	21.5	2.1	0.03	0.00	0.23	0.19	0.02		
france	2.62	4.45	-1.01	-0.77	-0.71	-0.01	0.02	19.0	11.8	10.5	0.0	0.0	0.23	0.13	0.11	0.00	0.00		
japon	7.74	0.85	0.40	0.20	-0.12	0.25	0.26	8.9	2.3	0.8	4.3	5.3	0.19	0.05	0.02	0.07	0.08		
CONTRIBUTION CUMULEE = 35.9 14.6 35.7 25.8 17.0 +-----																			
11 . GENRE																			
1	10.23	0.40	0.05	0.11	0.02	-0.06	-0.01	0.1	0.9	0.0	0.3	0.0	0.01	0.03	0.00	0.01	0.00		
2	1.93	6.39	-0.15	-0.36	0.06	0.21	0.08	0.3	1.9	0.1	0.8	0.1	0.00	0.02	0.00	0.01	0.00		
3	2.12	5.74	-0.08	-0.17	-0.14	0.11	-0.02	0.1	0.5	0.3	0.2	0.0	0.00	0.01	0.00	0.00	0.00		

----- CONTRIBUTION CUMULEE = 0.6 3.3 0.4 1.3 0.1 -----																	
12 . USAGE																	
1	9.73	0.47	0.08	0.06	0.01	-0.06	-0.02	0.5	0.3	0.0	0.4	0.0	0.02	0.01	0.00	0.01	0.00
2	2.81	4.09	0.55	-0.34	-0.26	-0.04	-0.54	5.9	2.5	1.5	0.0	8.0	0.07	0.03	0.02	0.00	0.07
3	1.00	13.31	-0.34	-0.52	-0.31	0.46	1.58	0.8	2.1	0.8	1.9	24.9	0.01	0.02	0.01	0.02	0.19
4	0.75	18.08	-2.68	1.14	1.29	0.35	0.13	38.3	7.4	9.7	0.8	0.1	0.40	0.07	0.09	0.01	0.00
----- CONTRIBUTION CUMULEE = 45.6 12.3 11.9 3.2 33.1 -----																	
13 . COEFB																	
0	7.67	0.86	-0.15	0.48	-0.52	-0.33	0.00	1.3	13.5	16.0	7.8	0.0	0.03	0.27	0.31	0.13	0.00
1	2.12	5.74	-0.36	-1.50	0.77	-0.31	-0.31	2.0	35.9	9.8	1.9	2.1	0.02	0.39	0.10	0.02	0.02
2	4.49	2.18	0.43	-0.12	0.52	0.72	0.16	6.0	0.5	9.4	21.2	1.1	0.09	0.01	0.12	0.24	0.01
----- CONTRIBUTION CUMULEE = 9.2 49.9 35.2 30.9 3.1 -----																	
14 . PUISS																	
1	4.93	1.90	-0.03	-0.03	-0.39	-0.14	-0.42	0.0	0.0	5.8	0.9	8.8	0.00	0.00	0.08	0.01	0.09
2	4.80	1.97	-0.24	-0.19	0.20	-0.01	0.58	1.9	1.3	1.5	0.0	16.0	0.03	0.02	0.02	0.00	0.17
3	4.55	2.14	0.28	0.23	0.21	0.16	-0.15	2.5	1.8	1.6	1.0	1.0	0.04	0.02	0.02	0.01	0.01
----- CONTRIBUTION CUMULEE = 4.5 3.1 8.9 1.9 25.8 -----																	
15 . DURC																	
a	11.04	0.29	0.02	0.08	0.09	-0.25	0.19	0.0	0.5	0.6	6.4	4.0	0.00	0.02	0.02	0.22	0.12
t	3.24	3.40	-0.05	-0.27	-0.29	0.86	-0.65	0.1	1.7	2.1	21.9	13.5	0.00	0.02	0.02	0.22	0.12
----- CONTRIBUTION CUMULEE = 0.1 2.2 2.8 28.3 17.4 -----																	

BIBLIOGRAPHIE

- [1] **A. Morineau, L. Lebart, M. Piron** ; Statistique exploratoire multidimensionnelle, DUNOD 1995
- [2] **Xavier GUYON et Michel NDOUMBE NKENG**, Cours de Modèle Linéaire et Extensions, Master de Statistique 2005 ; Université de Yaoundé I (Cameroun)
- [3] **GILBERT Saporta**, Probabilités, Analyse des données et Statistique 1990, Edition Technip 27 rue GINOUX 75737 Paris cedex 15
- [4] **Xavier BRY**, introduction à l'analyse factorielle des correspondances (simples et multiples) ; notes de cours ENSEA Abidjan (cote d'ivoire)
- [5] **C. HUBER**, Cours de modélisation Biostatistique en S-plus , Université Paris 5 , René Descartes UFR Biomédicale
- [6] **M. HALLING et J-F INGENBLEEK** (1978) Etude statistique des facteurs influençant le Risque automobile, la probabilité de sinistre Discussion paper n°5 Institut de Statistique de l'Université Libre de Bruxelles
- [7] **PAUL-MARIE Bernard**, Cours Régression Logistique, Université LAVAL Québec, CANADA
- [8] **Patrick. TAFFE**, cours de régression logistique appliquée, IUMSP Lausanne, Août 2004
- [9] **J.L. GOLMARD**, cours d'analyse de données censurées, Master de Statistique 2005 Université de Yaoundé (Cameroun)
- [10] **Didier D.-CASTELLE, Marie DUFLO**, Probabilités et Statistiques, tome1 : problèmes à temps fixe, MASSON, Paris, 1982
- [11] **Jean COURSOL**, Analyse de données et Datamining ; notes de cours, Master de statistique 2005, Université de Yaoundé (Cameroun)