PROJET  **et SAS**
ASSURANCE BELGES

Mots clés : Outils graphiques, Analyse des correspondances multiples, régression logistique, programmation légère

Les assurances belges cherchent à savoir ce qui discrimine les individus n'ayant pas eu de sinistre des individus ayant eu un sinistre ou plus.

Les données :

L'échantillon est constitué de 1106 assurés Belges observés en 1992 et répartis en 2 groupes.

- **les assurés qui n'ont eu aucun accident dans l'année**, qui sont au nombre de 556.
- **les assurés ayant eu au moins un sinistre dans** l'année; ils sont au nombre de 550.

On entend par assuré une personne physique ou morale.

Les variables de départ, concernent le souscripteur et le véhicule :

1 . Sinistralité RC - SNB11
2 . Code usage - CUSA (2 modalités)
3 . Age de l assuré - DNAI (9 modalités)
4 . Sexe - SEXE (3 modalités)
5 . Code Langue - CLAN (3 modalités)
6 . Code postal souscripteur - POSS2
7 . Bonus-malus RC(année en cours) - GBMC
8 . Puissance du véhicule (12 mod) - PUIS
9 . Bonus-malus Année -1 (2 mod) - GBM1
10 . Date effet Police (2 mod) - DPEP
11.Année de construction du véhicule (2 mod)
12 . Primes acquises RC 1991 en francs belges

Le but du projet est d'**utiliser ce jeu de données et de construire un modèle pour discriminer les assurés.**

Partie SAS

Étape préliminaire

I. Analyse exploratoire des données et analyse des correspondances multiples.

L'objectif de l'analyse exploratoire des données est de visualiser l'information contenue dans chacune des variables et d'évaluer l'impact des variables la présence ou non de sinistre.

1) Faire les analyses descriptives unidimensionnelles et des représentations graphiques pour synthétiser l'information contenue dans chaque variable.

2) Certaines modalités sont peu fréquentes et risquent de perturber l'analyse. Vous procéderez à des regroupements en 2 ou 3 modalités pour les variables concernées :

Age de l'assuré (3 mod) - DNAI: 1890-1949, 1950-73 et naissance inconnue
Code postal souscripteur (2 mod) - POSS2 : Bruxelles et autres
Bonus-malus Année en cours (2 mod) – GBMC : B-MC et Autres B-MC
Bonus-malus Année -1 (2 mod) – GBM1 : B-M 1 (-1) et Autres B-M (-1)
Date effet Police (2 mod) - DPEP 26-27 : avant 86 et autres
Puissance du véhicule (2 mod) - PUIS 32-33 : 10-39 Puis et 40-349 Puis
Année de construction du véhicule (2 mod) - DCOS 38-39 : 33-89 DCOS et 90-91 DCOS

3) Etudiez les liaisons entre les variables nominales transformées en 2) et la variable sinistre avec le chi de contingence. Observe-t-on des variables plus importantes dans la prédiction de variable sinistre *?*

4) Construire le tableau disjonctif contenant les indicatrices des variables précédentes que l'on notera avec les libellés suivants :

I_SINIS1
I_SINIS2
I_CUSAG1
I_CUSAG2
I_SEXE1
I_SEXE2
I_SEXE3
I_CLANG1
I_CLANG2
I_AGE3M1
I_AGE3M2
I_AGE3M3
I_CPOST1
I_CPOST2
I_DPOLI1
I_DPOLI2
I_BM_11
I_BM_12
I_PUIS1
I_PUIS2
I_DCONS1
I_DCONS2
I_PRIM1
I_PRIM2
I_PRIM3

5) Effectuer l'analyse des correspondances multiples (ACM) des variables nominales, la variable sinistre sera mise en supplémentaire. On exportera les coordonnées factorielles dans un fichier texte.

6) Faire une représentation du premier plan factoriel.

7) Les coordonnées des individus sur le premier axe factoriel peuvent être utilisée pour séparer les assurés ayant eu un sinistre des autres. Ceci constituera le modèle 0. On décide de classer les assurés à coordonnée positive en "PAS_SIN" et celles à coordonnée négative en "SIN". Générer la variable binaire correspondante.

8) En la comparant avec la variable sinistre calculez le pourcentage d'assurés mal classés ?

La question suivante est facultative

9) Interpréter le graphique. Le premier axe permet-il de discriminer significativement les assurés ayant eu un sinistre des autres ?

Partie R

II. Modèle de discrimination des assurés par la régression logistique

L'objectif de cette deuxième partie est d'une part de construire un modèle pour discriminer les assurés et d'évaluer la qualité du modèle à l'aide de Courbe ROC.

10) Importer le fichier texte créé en 5). Faire une représentation du premier plan factoriel.

11) Construire un modèle logistique en utilisant toutes les coordonnées factorielles issues de l'ACM. Cela constituera le modèle 1.

12) Calculer le pourcentage d'assurés mal classés. Comparez avec celui du modèle 0.
Représenter la courbe ROC associée à ce modèle. Donner l'AUC, la surface sous la courbe.