Master 1 Économétrie Statistiques

Projet Final : Analyse de données

Prutki Lucas
Barbey Charlotte
Domergue Corentin

1er février 2021

1 Régression binaire

1.1 Déterminer le meilleur modèle possible pour modéliser P(Y|X1, X2)

Après avoir importé les données, on transforme la variable Y en variable qualitative, cette variable est binaire et prend les valeurs 1,2. En première approche, on cherche le meilleur modèle en prédiction par les critères Bayesien AIC/BIC. Pour cela on part du modèle complet :

$$1 + X1 + X2 + X1 : X2$$

On y applique la fonction stepAIC avec les pénalités du critère AIC et BIC. Le modèle à AIC le plus faible est le modèle complet :

$$1 + X1 + X2 + X1 : X2$$

Celui à BIC le plus faible est un modèle réduit :

$$1 + X1 + X2$$

On cherchce le modèle le plus précis en prédiction, on regarde alors la matrice de confusion associée aux modèles avec les fonctions predict et table. A cette étape, on remarque que certains des modèles qu'on avait testé autres que ceux à AIC et BIC plus faibles sont meilleurs en prédiction que ces modèles. On pourrait l'expliquer par le fait qu'on a seulement 2 variables explicatives, on décide alors de tester tous les modèles possibles à partir du modèle complet :

$$1 + X1 + X2 + X1 : X2$$

$$1 + X2 + X1 : X2$$

$$1 + X1 + X1 : X2$$

$$1 + X1 + X2$$

$$1 + X1 : X2$$

$$1 + X2$$

$$1 + X1$$

Ainsi que tous ces mêmes modèles sans l'intercept (en remplaçant 1 par -1), soit un total de 14 modèles et on conclut avec les matrices de confusion que le modèle le plus prédictif est :

$$1 + X2 + X1 : X2$$

avec un taux de mauvaise prédiction de 28,6 % et donc une précision de 71,4%.

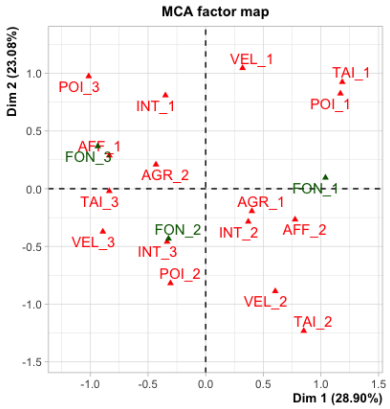
1.2 Utiliser le fichier "xsimutest" pour prédire 1000 observations de la variable Y du fichier de test. Il faut sauvegarder ces prédictions dans un fichier ".txt", les séparateurs devront être des retours chariot.

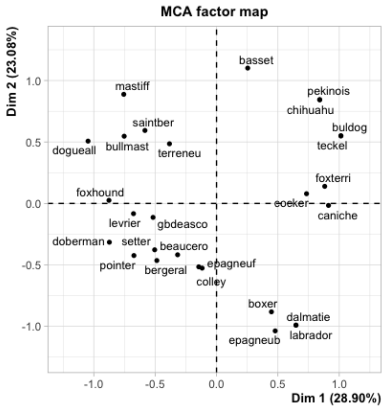
À partir du modèle choisi, on applique la fonction predict au fichier xsimutest et on code les prédictions de manière binaire, si la probabilité est > 0.5 on attribut 2 à la prédiction et sinon 1. On exporte ensuite le vecteur de prédictions dans un fichier txt. (cf. predictions.txt)

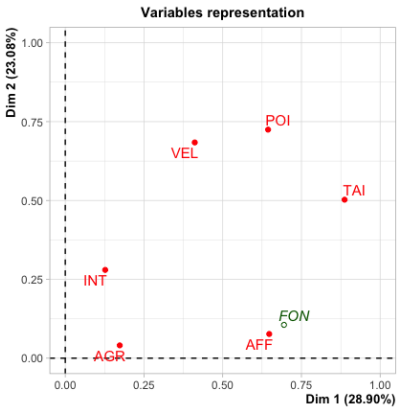
2 Race de chiens

2.1 En prenant la variable FON comme variable supplémentaire, faire une anayse des correspondances multiples de ces données.

On dispose d'un jeu de données « chiens » qui renseigne les caractéristiques de 27 races de chiens en utilisant sept variables. Les variables taille, poids, vélocité, intelligence et fonction ont trois modalités chacune. Les variables affectuosité et agressivité ont, quant à elles, deux modalités chacune. On fait une analyse des correspondances multiples (ACM) de ces données en mettant la variable fonction (FON) en variable supplémentaire. L'ACM est réalisée grâce à la commande suivante :
> res.mca=(chiens, quali.sup=7)





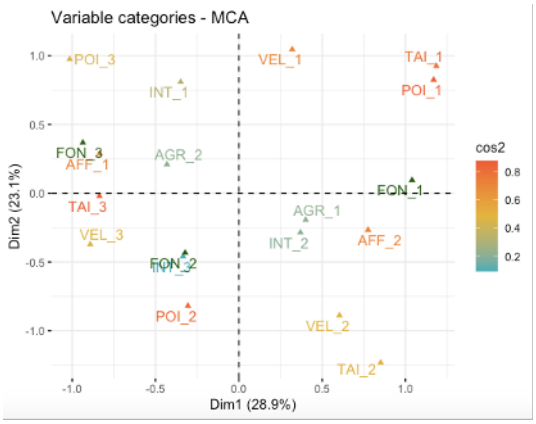


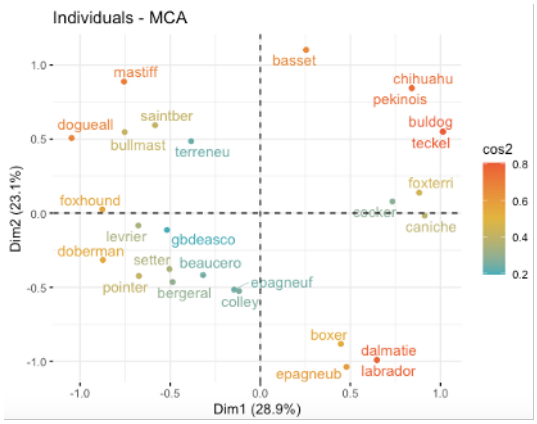
On affiche les valeurs propres :

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.481606165	28.896370	28.89637
Dim.2	0.384737288	23.084237	51.98061
Dim.3	0.210954049	12.657243	64.63785
Dim.4	0.157554025	9.453242	74.09109
Dim.5	0.150132670	9.007960	83.09905
Dim.6	0.123295308	7.397718	90.49677
Dim.7	0.081462460	4.887748	95.38452
Dim.8	0.045669757	2.740185	98.12470
Dim.9	0.023541911	1.412515	99.53722
Dim.10	0.007713034	0.462782	100.00000

Le pourcentage d'inertie expliqué par le premier plan factoriel est donc de 51,98 %.

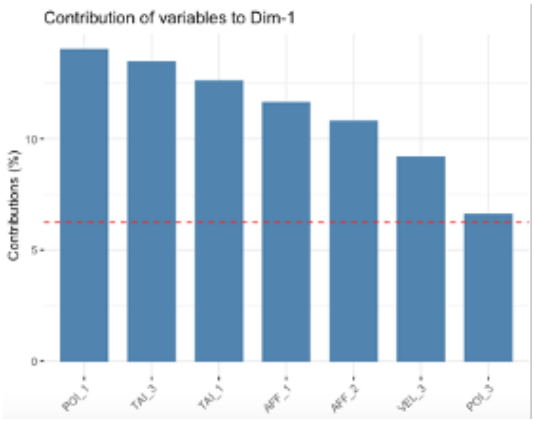
On donne la qualité de représentation des variables et des individus :

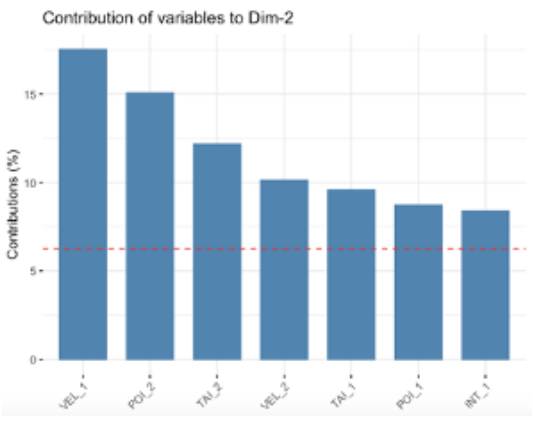


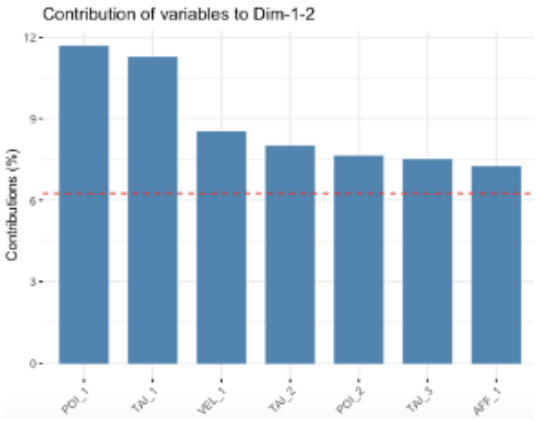


Le premier graphique nous montre que les catégories les moins bien représentées sont INT.3, INT.2, AGR.1, AGR.2 et INT.1 car elles ont des cos2 compris entre 9% et 33%. Ainsi, l'intelligence et l'agressivité des chiens ne sont pas bien prises en compte dans ces données. Le second graphique montre que les individus les moins bien représentés sont Terreneu, Gbdeasco, Beaucero, Bergeral, Colley et Cooker car ils ont des cos2 inférieurs à 40%.

On donne la contribution des variables respectivement aux axes 1, 2 et 1-2 :

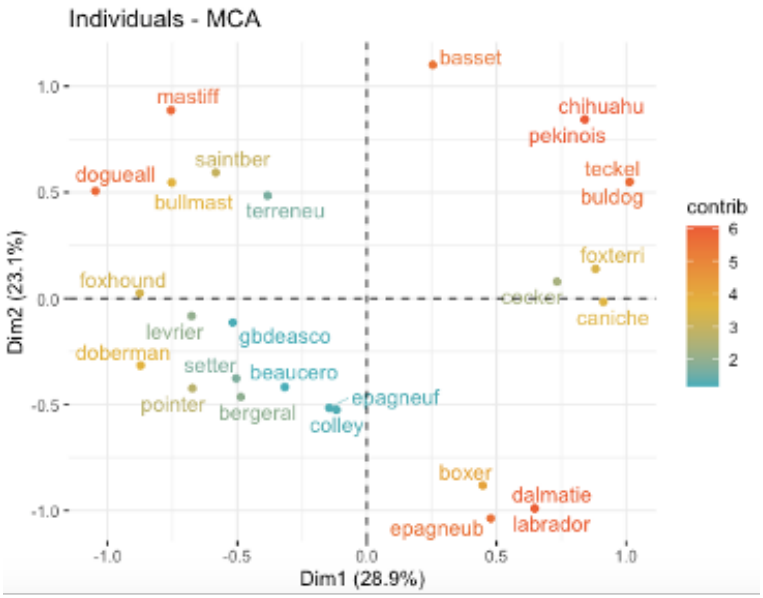






On constate que le premier axe est déterminé par les chiens lourds (POI.3) et grands (TAI.3) d'un côté, et ces derniers s'opposent aux chiens légers (POI.1) et petits (TAI.1), de l'autre. L'axe 1 est principalement déterminé par le poids du chien. Le deuxième axe est, quant à lui, déterminé par les chiens lents (VEL.1) et lourds (POI.3) ou petits (TAI.1) d'un côté, et ces derniers s'opposent aux chiens moyennement grands (TAI.2) et rapides (VEL.2). L'axe 2 est principalement défini par la vélocité du chien. Les variables citées pour interpréter les axes sont bien représentées.

On donne la contribution des individus :



Les chiens qui contribuent le plus à l'axe 1 sont Buldog et Teckel qui s'opposent à Dogueall. Ceux qui contribuent le plus à l'axe 2 sont Basset et Mastiff qui s'opposent à Epagneub, Dalmatie, Labrador et Boxer.

2.2 En déduire une description des différentes races de chiens.

Ainsi, on en déduit les caractéristiques suivantes des races de chiens, on interprète seulement les races de chiens bien représentées sur le premier plan factoriel :

- Buldog, Teckel, Chihuahu et Pekinois : petits et légers
- Dogueall : grand et lourd
- Foxhoun et Doberman : grands et très rapides
- Basset : lent et poids moyen
- Mastiff : lourd et lent
- Epagneub, Dalmatie, Labrador et Boxer : moyens et rapides (vélocité moyenne)

Remarque :
En observant, le premier graphique obtenu en faisant la ACM, on peut en déduire que FON.3 est liée à la partie gauche de l'axe 1 et que FON.1 à sa partie droite. Aisni, les chiens bien représentés et présents sur la partie droite de l'axe 1 seraient des chiens de compagnie, ceux situés à gauche de l'axe 1 seraient potentiellement des chiens de garde.