Université d'Angers

master mathematiques et applications

M1 DATA SCIENCE

Année académique 2019-2020

TRAVAIL ENCADRÉ DE RECHERCHE

Techniques de scoring et applications

Etudiantes :
Nadia GHERNAOUT
Philippine RENAUDIN

Tuteur Enseignant :
Frédéric Proïa

22 mai 2020

Projet

Nadia GHERNAOUT
Philippine RENAUDIN

Table des matières

1 Introduction générale

L'idée générale du scoring est d'affecter une note (un score) globale à un individu à partir de plusieurs descripteurs, quantitatifs ou qualitatifs. À partir de cette note, on affecte l'individu à un groupe préexistant. Un score peut donc être définit comme un outil statistique ou probabiliste de détection de risque. Le scoring peut également être vu comme l'application au monde de l'entreprise de plusieurs techniques de classement. Nous en aborderons 2 dans ce rapport.

Nous pouvons déjà citer plusieurs types de score :

1. Les scores de risque :
   — risque de crédit ou credit scoring : prédire le retard de remboursement de crédit.
   — risque financier : prédire la bonne ou mauvaise santé d'une entreprise.
   — risque médical : prédire l'apparition d'une maladie chez un patient.
2. Les scores en marketing :
   — score d'attrition : prédire le risque qu'un client passe à la concurrence ou résilie son abonnement.
   — score d'appétence : prédire l'appétence d'un client à acheter tel ou tel type de produit.

La création d'un score se fait en fonction des objectifs recherchés et des moyens techniques disponibles. Par exemple, le développement d'un score comportemental nécessite de disposer de données sur au moins un an, si l'on a moins d'historique, il vaut mieux partir sur un score générique ou un score d'octroi.

Il faut aussi également choisir l'utilisation qui sera faite du score : outil d'aide à la décision ou outil de ciblage pour le marketing direct par exemple. C'est en fonction de l'utilisation que l'on en fera que la règle de décision sera ajustée.

Pour construire un score, il faut dans un premier temps disposer d'un échantillon suffisamment conséquent pour pouvoir tester plusieurs modèles prédictifs. De plus, pour éviter des problèmes de surestimation de la qualité du modèle, il est préférable de séparer l'échantillon d'étude en deux sous-échantillons : un échantillon d'apprentissage à partir duquel sera créé le modèle, et un échantillon test sur lequel sera testé la qualité du modèle par rapport à l'objectif recherché et au risque que l'on est prêt à prendre.

Ensuite, il faut élaborer un modèle prédictif à l'aide de techniques prédictives : analyse discriminante et régression logistique en l'occurrence.

Enfin, les notes de score sont découpées en plusieurs classes de valeur. Dans le domaine financier, on aura tendance à découper les notes de score en trois classes : faible, moyen, fortes. Dans le milieu médical, on préfèrera 2 classes : à risque, non à risque. La règle de classement (seuil comparatif du score) se décide en fonction du risque d'erreur que l'on souhaite prendre.

Nous présentons dans ce rapport 2 des techniques prédictives les plus utilisées en scoring : la régression logistique et l'analyse discriminante. Pour illustrer ce qu'est le scoring, nous avons utilisé ces 2 techniques sur 2 jeux de données différents. Nous présentons dans la suite la théorie de chaque technique ainsi que l'étude des données associée.

Première partie

Protocole

1 Validation croisée

Comme il a été mentionné en introduction, une habitude à prendre lors de toute analyse de données est de séparer les données en plusieurs sous échantillons pour éviter les problèmes de surestimation des capacités du modèle.

On appelle validation croisée la technique consistant à ajuster un modèle prédictif sur un échantillon d'apprentissage et à valider ce modèle sur un échantillon test. Ces échantillons peuvent provenir du même jeu de données auquel cas il est coutume que l'échantillon d'apprentissage représente entre 60% et 80% des données et que l'échantillon test représente 20% à 40%. Il est également possible, si l'on dispose de plusieurs jeux de données différents pour le sujet d'étude, de prendre un jeu de données comme échantillon d'apprentissage et de valider le modèle sur un deuxième jeu de données.

Nous avons choisi dans nos études de cas de réaliser une validation croisée à partir d'un seul jeu de données en le décomposant en un échantillon d'apprentissage représentant 80% du jeu de données et en un échantillon test représentant les 20% restant, et ce de manière aléatoire.

2 Sélection de modèles

Une fois le jeu de données séparé en deux échantillons, vient le moment de construire différents modèles prédictifs. Cependant, il n'est pas toujours évident de savoir quelles variables garder, quelles sont celles qui apportent le plus d'information, qui discriminent le mieux les groupes d'individus, etc... C'est pourquoi on s'appuie sur différents indicateurs, en plus de ceux implémentés par défaut dans les logiciels. Nous en évoquons 3 ici :

2.1 Tests entre modèles emboîtés

A l'image de ce qui est fait en régression linéaire il existe des tests entre modèles emboités, on souhaite comparer un modèle restreint de p_0 paramètres au modèle global (à p paramètres).

Soit $p_0 < p$, on compare le modèle $\mathcal{M}_0$

$$\text{logit}(p_\gamma(x)) = \gamma_1 x_1 + \cdots + \gamma_{p_0} x_{p_0}$$

avec le modèle $\mathcal{M}_1$

$$\text{logit}(p_\beta(x)) = \beta_1 x_1 + \cdots + \beta_p x_p$$

On peut ainsi faire le test d'hypothèses suivant :

$$\mathcal{H}_0 : "\beta_{p_0+1} = \cdots = \beta_p = 0" \text{ contre } \mathcal{H}_1 : "\exists j \in \{p_0 + 1, \cdots, p\} : \beta_j \neq 0"$$

Ne pas rejeter $\mathcal{H}_0$ signifie privilégier le modèle $\mathcal{M}_0$ au détriment du modèle $\mathcal{M}_1$.

Ce test peut être réalisé à l'aide du test de Wald ou test de rapport de vraisemblance(déviance).

2.1.1 Le test de Wald

2.1.2 La déviance

Nous verrons plus tard dans ce rapport que les paramètres des différents modèles sont le plus souvent estimés par la méthode du maximum de vraisemblance.

Pour savoir quels modèles garder, il est donc courant d'utiliser le critère de la déviance. En effet, la déviance est égale à -2log-vraisemblance donc le modèle maximisant la vraisemblance est celui minimisant la déviance. On utilise plutôt cet indicateur car plus simple à calculer que les expressions du maximum de vraisemblance.

$$Deviance = -2\ln\left(\frac{\text{Vraisemblance sans la variable}}{\text{Vraisemblance avec la variable}}\right)$$

On a sous $\mathcal{H}_0$:

$$2\left(\ell\ell_{\widehat{\beta}} - \ell\ell_{\widehat{\beta^0}}\right) \longrightarrow \chi^2(p)$$

Avec $\ell\ell_{\widehat{\beta}}$ la log-vraisemblance du modèle avec la variable, et $\ell\ell_{\widehat{\beta^0}}$ la log vraisemblance du modèle sans la variable.

Il existe des fonctions dans Rstudio rendant le /les modèles minimisant la déviance, il n'est donc pas nécessaire de créer les modèles au préalable et de les comparer entre eux, le logiciel fait ce travail à notre place.

2.2 Méthodes automatiques : critères AIC et BIC

L'approche du test de Wald et test du rapport de vraisemblance permet de choisir un modèle parmi deux modèles emboités. Cependant ces tests ne permettent pas de sélectionner automatiquement un sous groupe de variables explicatives.

Pour des modèles ayant un nombre de paramètres égal, l'algorithme utilisera la vraisemblance pour choisir le meilleur modèle à k variables. Cependant la vraisemblance ne pourra pas être utilisée quand le nombre de paramètres sera différent pour des modèles. En effet la vraisemblance augmente avec le nombre de paramètres, ainsi le modèle choisi sera celui avec le plus grand nombre de paramètres.

Pour pallier à cela des critères existent. Parmi les critères les plus utilisés, on retrouve, comme pour les modèles linéaires l'AIC et le BIC. Ces critères pénalisent l'opposé de la log-vraisemblance d'un modèle $\mathcal{M}$ par son nombre de paramètres k.

AIC (*Akaike Information criterion*)

$$AIC = -2\ell\ell(\widehat{\beta}) + 2k$$

BIC (*Bayesian information criterion*)

$$BIC = -2\ell\ell(\widehat{\beta}) + \ln(n)k$$

avec $\ell\ell(\widehat{\beta})$ qui désigne la log-vraisemblance maximisée du modèle $\mathcal{M}$, ces critères sont basés sur deux parties :

- la composante $-2\ell\ell(\widehat{\beta})$ mesure l'ajustement du modèle aux données. Plus les valeurs sont faibles plus l'ajustement est bon.
- les composantes $2k$ pour l'AIC et $k\ln(n)$ pour le BIC mesurent la complexité du modèle

Les modèles qui réalisent un bon compromis entre qualité d'ajustement et complexité, correspondront aux modèles minimisant le BIC ou l'AIC.

Remarque 1. Le critère BIC aura tendance à choisir des modèles plus parcimonieux que le critère AIC. Cela arrive quand $\ln(n) > 2$ soit dès que le modèle a 8 observations ou plus.

Remarque 2. Les fonctions R correspondant à ces deux critères sont :

- `bestglm` : qui utilise l'algorithme de *Best subset selection* qui nous laisse le choix de sélectionner le critère à utiliser pour trouver le meilleur modèle (AIC, BIC, EBIC, CV pour la validation croisée, $\cdots$)
- `step` : fonction qui trouve le modèle minimisant l'AIC. Elle permet de lancer les procédures pas à pas. En effet l'algorithme nécessitant le calcul de 2^p modèles devient coûteux en temps de calcul lorsque le nombre de variables p est grand (au delà de 30).

2.3 Le lambda de Wilks

Cet indicateur est propre à l'analyse discriminante et n'est pas utilisé en régression logistique. Le Lambda de Wilks est souvent utilisé dans les logiciels comme critère pour ne garder que les variables apportant de l'information sur l'appartenance ou non d'un individu à un groupe.

Le Lambda de Wilks est une approche paramétrique permettant de tester si plusieurs variables continues distinctes $X = (X_1, \cdots, X_p)$ sont liées à une variable qualitative Y à $K \geq 2$ groupes, lorsqu'elles sont considérées avec leurs différentes interactions multivariées.

Les hypothèses d'utilisation de ce test sont : $X_{|Y=1}, \cdots, X_{|Y=k}$ suivent une loi normale et leur matrice de covariance respective sont égales (homoscédasticité).

La statistique du test du Lambda de Wilks se définie de la manière suivante :

$$\Lambda = \frac{\det(W)}{\det(B)}$$

Où W est la matrice de variance-covariance intragroupe et B la matrice de variance-covariance intergroupe.

Cette statistique de test suit une loi de Wilks à $(P, n, K-1)$ degrés de liberté et l'hypothèse H_0 est : « Indépendance entre X et $Y|_{\mu_1=\cdots=\mu_k}$ ».

Une variable a un bon pouvoir discriminant si la dispersion intra-groupe est faible et si la dispersion intergroupe est forte. Donc plus le Lambda de Wilks sera faible, plus la variable considérée est discriminante. C'est ce critère qu'utilise la commande `greedy.wilks` de Rstudio que nous avons utilisée pour trouver les variables les plus discriminantes dans notre jeu de données et ainsi se focaliser sur un nombre de modèles plus réduit.

3 Qualité du modèle

3.1 Matrices de confusion

L'analyse factorielle discriminante probabiliste et la régression logistique sont utilisées pour affecter des scores à des individus. Mais donner un score à un individu sans contexte d'étude est assez absurde. C'est pourquoi, il faut évaluer le modèle choisi pour savoir si celui-ci permet une bonne prédiction des groupes d'appartenance.

Il est coutume de tester la qualité du modèle grâce aux matrices de confusions qui donnent le taux de bon et mauvais classement des individus dans chaque groupe. On construit donc souvent plusieurs modèles et évaluons leurs capacités prédictives grâce à ces matrices.

Dans le cas de deux groupes les matrices de confusions se représentent souvent de la façon suivante :

	$Y = 0$	$Y = 1$
$Y_{pred} = 0$	VN	FN
$Y_{pred} = 1$	FP	VP

Avec :

— 0 = négatif

— 1 = positif

— VN = vrai négatif

— FP = faux positif

— FN = faux négatif

— VP = vrai positif

En fonction de ce que l'on cherche à faire grâce au scoring, on préfèrera retenir le modèle minimisant les faux négatifs ou les faux positifs, ou encore le modèle maximisant les vrais positifs ou vrais négatifs.

3.2 Courbes ROC et AUC

Une fois notre modèle choisi grâce aux comparaisons des différentes matrices de confusion, il est possible de visualiser graphiquement la qualité globale de ce modèle grâce à une courbe ROC.

Courbe ROC Une courbe ROC (*Receiver Operating Curve*) est un graphique représentant les performances d'un modèle de classification pour tous les seuils de classification. Cette courbe trace le taux de vrais positifs en fonction du taux de faux positifs :

— Le taux de vrais positifs (TVP) (sensibilité) est défini comme : $TVP = \frac{VP}{VP+FN}$

— Le taux de faux positifs (TFP) (spécificité) est défini comme : $TFP = \frac{VN}{VN+FP}$

.

Les termes "positifs" et "négatifs" dépendent de ce que l'on aura choisi au préalable.

La *sensibilité* se définit comme le pourcentage de vrais positifs : $1-\beta$: D'un point de vue médical cela veut dire être testé positif à un test détectant la présence de maladie quand on est bien malade .

La *spécificité* se définit quant à elle comme le pourcentage de vrais négatifs : $1-\alpha$. D'un point de vue médical cela signifie être testé négatif à un test détectant la présence de maladie, quand on est bien sain.

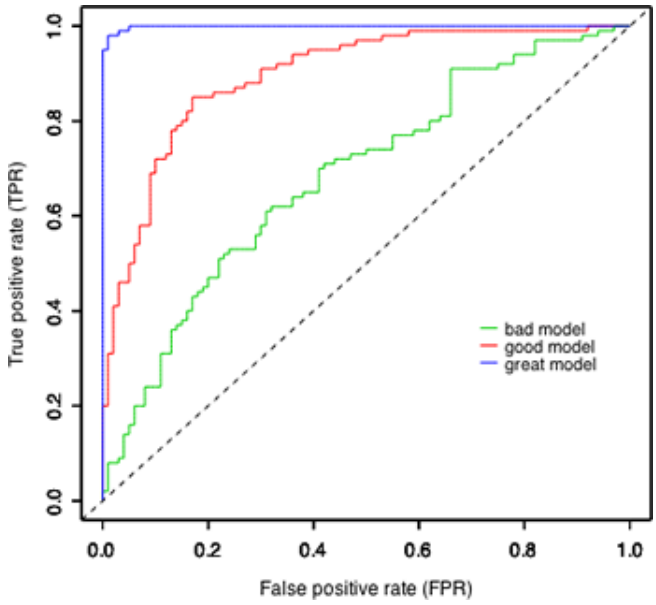


FIGURE 1 – Exemple de courbes ROC (Source : *The University of North California at Chapell Hill*)

Ainsi la courbe ROC représente le taux de vrais positifs (TVP) par rapport au taux de faux positifs (TFP). Ainsi elle représente la sensibilité sur 1− spécificité. Sur ces représentations de courbes ROC :

— La courbe verte caractérise un mauvais modèle

— La courbe rouge caractérise un bon modèle

— La courbe bleue caractérise un excellent modèle

Plus la courbe est proche du coin supérieur gauche, meilleur est le modèle. Si la courbe ROC correspond à la diagonale en pointillés cela veut dire que

AUC Représente l'aire sous la courbe ROC, elle vaut au maximum 1 lorsque le modèle est parfait.

3.3 Histogrammes de scores

Le seuil utilisé dans l'algorithme de recherche du meilleur modèle (pour la problématique considérée) est par défaut fixé à 0.5 (dans le cas de 2 groupes) pour décider du groupe d'appartenance. Or toute l'essence du scoring est justement de trouver le seuil qui donnera un risque d'erreur le plus faible possible, tout en prenant les contraintes de coût financier en compte.

C'est pourquoi il est courant de représenter les histogrammes de score pour déterminer le seuil optimal.

On trace les histogrammes des scores des individus en fonction de leur vrai groupe d'appartenance. Dans le cas où les 2 histogrammes sont disjoints, alors il est possible de trouver un seuil qui annulera l'erreur prise, mais ce cas est plutôt rare. Les histogrammes sont toujours plus ont moins superposés et c'est donc le travail de l'analyste de choisir le seuil qui minimisera le plus possible le taux d'erreur pris, tout en prenant en compte encore une fois toutes les contraintes financières ou matérielles.

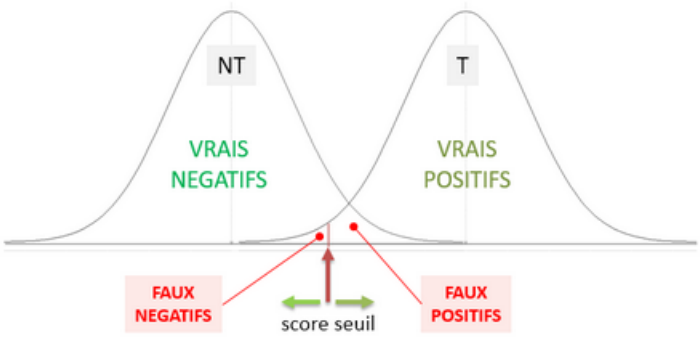


FIGURE 2 – Exemple d'histogrammes de scores (Source : Cours de Psychométrie de Mr Jean-Luc Roulin)

Deuxième partie

Analyse Factorielle Discriminante

1 Théorie de la méthode probabiliste

Le but premier de cette méthode est de prédire au mieux les valeurs d'une variable Y qualitative à K modalités, à partir de p variables explicatives $X = (X_1, \cdots, X_p)$ quantitatives.

Dans cette section, nous allons classer un individu dans le groupe le plus "probable" et non dans le groupe le plus "proche" comme c'est le cas pour l'analyse discriminante géométrique. Pour ce faire nous allons devoir faire des hypothèses probabilistes sur les données.

On suppose maintenant que les données sont issues d'une population regroupant des individus de K groupes prédéfinis différents $G_1, \cdots, G_k$ et que :

— Y est une variable aléatoire qui prend ses valeurs dans $\{1, \cdots, K\}$

— $X = (X_1, \cdots, X_p)$ est un vecteur de variables aléatoires réelles

On notera :

— $I(k)$ représente l'ensemble des indices i des n_k individus du groupe G_k.

— $p_k = \mathbb{P}(Y = k)$ la probabilité à priori d'appartenir au groupe G_k.

— $f_k = \mathbb{R}^p \longrightarrow [0, 1]$ la densité de X dans le groupe G_k

— w_i le poids d'un individu x_i

— $w_k = \dfrac{n_k}{n}$ le poids du groupe G_k où n_k représente le nombre d'individus dans le groupe G_k.

— $g_k = \displaystyle\sum_{i \in I(k)} \frac{w_i}{w_k} x_i$ le centre de gravité du sous nuage formé par les individus du groupe G_k

— $W_k = \displaystyle\sum_{i \in I(k)} \frac{w_i}{w_k} (X_i - g_k)(X_i - g_k)^T$ la matrice de variance covariance du nuage N_k représentant le groupe G_k

— $W = \displaystyle\sum_{k=1}^{q} w_k W_k$ la matrice de variance covariance intra-classes

On supposera que l'on dispose d'un échantillon i.i.d $(X_1, Y_1), \cdots, (X_n, Y_n)$ de même loi que (X, Y).

1.1 Approche probabiliste : modèle Bayésien

Avec les notations précédentes, on obtient d'après la formule de Bayes que la probabilité qu'un individu x appartienne au groupe G_k s'écrit :

$$\mathbb{P}(G_k \mid x) = \frac{p_k f_k(x \mid k)}{\sum_{u=1}^{K} p_u f_u(x \mid u)}$$

On affecte alors l'individu x au groupe G_k pour lequel la probabilité $\mathbb{P}(G_k \mid x)$ est la plus forte (c'est la règle du minimum a posteriori).

Or comme le dénominateur est constant pour un individu x donné, il suffit de déterminer le groupe pour lequel $p_k f_k(x \mid k)$ est le plus grand.

La fonction de densité f_k introduite ici peut être déterminée soit :

1. Par des approches non paramétriques : on cherche à estimer directement à partir des données les densités (méthode des noyaux, des plus proches voisins).
2. Par des méthodes paramétriques : on suppose $f_k(x)$ d'une forme paramétrique particulière et on estime les paramètres grâce à l'échantillon d'apprentissage.

1.2 Modèle Bayésien avec méthode paramétrique

On considère dans cette section que X suit une loi normale $\mathcal{N}(\mu_k; \Sigma_k)$ dans chaque groupe G_k. On se place donc dans le cas paramétrique Gaussien.

On a donc

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} (\det(\Sigma_k))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

où μ_k appartient à $\mathbb{R}^p$ est le vecteur des moyennes théoriques et Σ_k la matrice des variances-covariances théoriques. La règle de classement énoncée précédemment (également appelée règle de Bayes) revient donc à maximiser

$$p_k \times \frac{1}{(2\pi)^{\frac{p}{2}} (\det(\Sigma_k))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right)$$

Par soucis pratique, on préfère maximiser le logarithme de cette expression, c'est-à-dire maximiser :

$$\begin{aligned}\ln(p_k f_k(x)) &= \ln(p_k) + \ln(f_k(x)) \\ &= \ln(p_k) - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln(\det(\Sigma_k)) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\end{aligned}$$

Cela revient à maximiser

$$\begin{aligned}&-(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - p\ln(2\pi) - \ln(\det(\Sigma_k)) + 2\ln(p_k) \\ =& -x^T\Sigma_k^{-1}x + \underbrace{x^T\Sigma_k^{-1}\mu_k + \mu_k^T\Sigma_k^{-1}x}_{=2\mu_k^T\Sigma^{-1}x \text{ car symétrique}} - \mu_k^T\Sigma_k^{-1}\mu_k - p\ln(2\pi) - \ln(\det(\Sigma_k)) + 2\ln(p_k)\end{aligned}$$

1.2.1 Analyse discriminante linéaire

Dans le cas où les matrices de variance- covariance Σ_k peuvent être supposées égales, on obtient le critère linéaire en x (d'affectation de l'individu x au groupe G_k) :

$$SL_k(x) = 2\mu_k^T\Sigma^{-1}x - \mu_k^T\Sigma^{-1}\mu_k + 2\ln(p_k)$$

car $\det(\Sigma)$ et $x^T\Sigma^{-1}x$ est constant pour un individu x donné.
On affecte donc x au groupe G_k qui donne une valeur de $SL_k(x)$ maximale.

1.2.2 Analyse discriminante quadratique

Dans le cas où les matrices de variance-covariance Σ_k ne peuvent pas être supposées égales, on obtient le critère quadratique en x.

$$SQ_k(x) = -x^T\Sigma_k^{-1}x + 2\mu_k^T\Sigma_k^{-1}\mu_k - \ln(\det(\Sigma_k)) + 2\ln(p_k)$$

On affecte donc x au groupe G_k qui donne une valeur de $SQ_k(x)$ maximale.

1.3 Estimation des paramètres

Le calcul de $SL_k(x)$ et $SQ_k(x)$ n'est possible qu'en estimant les paramètres μ_k, p_k et Σ_k non connus en pratique. Ces paramètres peuvent être estimés par maximum de vraisemblance auquel cas on obtient :

$$\begin{aligned}
\widehat{\mu_k} &= g_k \\
\widehat{p_k} &= \frac{n_k}{n} = w_k \\
\widehat{\Sigma_k} &= \begin{cases} \widehat{\Sigma} = \dfrac{n}{n-K}W & \text{(cas homoscédastique)} \\ \widehat{\Sigma_k} = \dfrac{n_k}{n_k-1}W_k & \text{(cas hétéroscédastique)} \end{cases}
\end{aligned}$$

1.3.1 Analyse discriminante linéaire (LDA)

Dans le cas homoscédastique, on a alors :

$$L_k(x) = 2g_k^T\widehat{\Sigma}^{-1}x - g_k^T\widehat{\Sigma}^{-1}g_k + 2\ln(\widehat{p_k})$$

Cette fonction est appelée fonction linéaire discriminante du groupe G_k. Chaque fonction linéaire discriminante définie une fonction de score et un nouvel individu sera donc affecté au groupe G_k pour lequel le score sela le plus élevé.

Remarque 3. On retrouve la fonction linéaire discriminante

$$L_k(x) = x^TW^{-1}g_k - \frac{1}{2}g_k^TW^{-1}g_k$$

de l'analyse discriminante géométrique avec le terme $\ln(p_k)$ en plus. Dans le cas où l'on fait l'hypothèse d'égalité des probabilités à priori ($p_1 = \cdots = p_K$), la règle de l'analyse discriminante linéaire (LDA) est équivalente à la méthode de l'analyse discriminante géométrique.

1.3.2 Analyse discriminante quadratique (QDA)

Dans le cas hétéroscédastique, on a alors :

$$Q_k(x) = -(x-g_k)^T\widehat{\Sigma}^{-1}(x-g_k) - \ln(\det(\widehat{\Sigma_k})) + 2\ln(\widehat{p_k})$$

Cette fonction est appelée fonction quadratique discriminante du groupe G_k. Chaque fonction quadratique discriminante définie une fonction score et un nouvel individu sera donc affecté au groupe G_k pour lequel le score sera le plus élevé.

1.4 Cas particulier de 2 groupes

On se place dans le cas où $K = 2$ et $\Sigma_1 = \Sigma_2 = \Sigma$ (cas homoscédastique).
On affecte donc un individu x au groupe 1 si $p_1 f_1(x; \mu_1, \Sigma) > p_2 f_2(x; \mu_2, \Sigma)$. Ce qui donne après calculs :

$$x^T \Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) + \ln\left(\frac{p_2}{p_1}\right)$$

On pose donc

$$S(x) = x^T \Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) - \ln\left(\frac{p_2}{p_1}\right)$$

La nouvelle fonction discriminante est

$$L(x) = x^T \widehat{\Sigma}^{-1}(g_1 - g_2) - \frac{1}{2}(g_1 + g_2)^T \widehat{\Sigma}^{-1}(g_1 - g_2) - \ln\left(\frac{\widehat{p_2}}{\widehat{p_1}}\right)$$

On compare donc cette fonction à 0 pour affecteur x dans un des deux groupes :

— si $L(x) \geq 0 \Longrightarrow x$ est affecté au groupe 1

— si $L(x) < 0 \Longrightarrow x$ est affecté au groupe 2

Dans ce cas de figure précis, on a

$$\mathbb{P}(G_1 \mid x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)} = \frac{1}{1 + \frac{p_2 f_2(x)}{p_1 f_1(x)}} = \frac{1}{1 + e^{-S(x)}}$$

Donc la probabilité à posteriori que $x \in G_1$ est la fonction logistique de $S(x)$.

2 ACP préliminaire

2.1 Présentation du jeu de données

Notre jeu de données traite de la situation économique d'exploitations fermières entre 1988 et 1994 dans 3 communes Françaises.

Le contexte de l'époque se caractérisait par des réformes politiques en agriculture, l'agrandissement de l'Union Européenne mais aussi des durcissements des contraintes économiques dans le secteur agricole dans les années 80. Tout cela a eu un impact sur les exploitations fermières françaises et a résulté en une multiplication des crises agricoles. C'est à partir de ce moment que s'est posée la question d'estimer les risques financiers en agriculture.

L'expérience a démontré que les mesures de redressement financier pouvaient être efficaces pourvu que des actions préventives débutent tôt. Il est donc important d'avoir une méthode pour la détection rapide des risques financiers en agriculture.

La distinction entre ferme défaillante et ferme saine est basée sur le concept de viabilité et d'insolvabilité des exploitations fermières.

L' *insolvabilité* est définie comme la situation dans laquelle une exploitation agricole n'est pas en mesure d'honorer les obligations générées par la dette existante, à savoir le paiement des intérêts et le paiement des prêts. Ainsi appliquer une méthode de « credit scoring » permet de diagnostiquer de manière préventive les problèmes financiers des exploitations fermières.

Plusieurs critères contribueraient à la déstabilisation des fermes d'un point de vue financier :

- le déclin des prix des produits agricoles
- l'augmentation des crédits
- affaiblissement financier dû à :
  - une augmentation des dépenses
  - une baisse du chiffre d'affaire
  - une recrudescence des incidents et des retards de paiement

Les données concernent 1260 exploitations fermières réparties en 2 groupes (décrits par la variable DIFF). Le premier groupe rassemble les fermes saines (653) et le second groupe rassemble les fermes défaillantes (607). La variable à expliquer est donc la variable DIFF.

Plusieurs variables explicatives sont présentes :

- `CNTY` : code de département
- `DIFF` : variable à expliquer, est ce que la ferme a déjà eu un incident de paiement (1= ferme saine, 2= ferme défaillante)
- `STATUS` : statut légal (1 = propriétaire indépendant, 0= entreprise)
- `HECTARE:` aire de la ferme en hectares
- `ToF` : index de type de ferme
- `OWNLAND` : owned land (O= Oui, N= non)
- `AGE` : l'âge du propriétaire des terres
- `HARVEST` : année de récolte concernée

De plus, pour calculer les risques financiers plusieurs ratios quantitatifs sont présents dans le jeu de données. Ces ratios définissent un ensemble de critères micro économiques qui quantifient le degré de faillite des exploitations fermières.

Nous choisissons de prendre ces ratios comme variables explicatives actives pour la suite de l'étude, et non celles mentionnées précédemment.

- Capitalisation : R1, R2, R3, R4, R5
- Poids des dettes : R6,R7,R8
- Liquidité :R11, R12, R14 mesure la capacité d'une entreprise à s'acquitter de ses dettes à court terme.
- Debt servicing : R17, R18, R19, R21, R22 :) mesure la capacité d'une entreprise à utiliser leurs bénéfices pour rembourser toutes ses dettes de court et long termes.
- Capital Profitability : R24
- Earnings : R28, R30, R32 (gains, bénéfices, profits, revenus)
- Productive activity : R36, R37

La première méthode nous permettant de construire une score est l'Analyse en Composantes Principales (ACP). Nous allons voir qu'on peut faire du scoring avec ce genre de méthodes.

2.2 Analyse en composantes principales

Analyse exploratoire des données :

	CNTY	DIFF	STATUS	HECTARE	TOF	OWNLAND	AGE	HARVEST	R1	R2	R3	R4	R5	R6	R7	R8
1	27	1	1	166	1	1	35	88	0.449	0.622	0.2550	0.11450	0.334	0.785	0.585	0.2002
2	27	1	1	101	1	0	35	88	0.450	0.617	0.2411	0.10840	0.341	0.518	0.393	0.1250
3	27	1	1	138	3	1	42	88	0.332	0.819	0.5568	0.18510	0.147	0.700	0.310	0.3895
4	27	1	1	166	1	1	50	88	0.363	0.733	0.3596	0.13050	0.232	0.773	0.495	0.2779
5	27	1	0	137	4	1	33	88	0.440	0.650	0.3142	0.13820	0.302	0.846	0.580	0.2658
6	27	1	1	107	1	1	45	88	0.306	0.755	0.2635	0.08055	0.225	0.709	0.523	0.1870

	R11	R12	R14	R17	R18	R19	R21	R22	R24	R28	R30	R32	R36	R37
1	0.6628	1.3698	0.2320	0.0884	0.0694	0.1660	0.1340	0.3219	0.295	0.475	0.3500	0.4313	0.886	0.572
2	0.7098	1.2534	0.1497	0.0671	0.0348	0.1360	0.0802	0.3133	0.365	0.434	0.2978	0.3989	0.351	0.867
3	0.4142	0.6370	0.4847	0.0445	0.0311	0.1030	0.0890	0.2945	0.166	0.350	0.2468	0.3187	1.300	0.475
4	0.4661	1.0698	0.3735	0.0621	0.0480	0.1080	0.0851	0.1909	0.265	0.475	0.3759	0.4313	1.385	0.470
5	0.7715	1.4752	0.2563	0.0489	0.0414	0.1270	0.0838	0.2567	0.257	0.475	0.3669	0.4313	0.886	0.520
6	0.8178	1.4682	0.1861	0.0243	0.0173	0.0652	0.0390	0.1472	0.191	0.443	0.3759	0.4257	1.316	0.431

FIGURE 3 – Visualisation des données pour les 6 premières fermes

Nous explorons quelques ratios pour visualiser leurs potentiels effets sur la variable à expliquer `DIFF`. En bleu sont représentées les fermes défaillantes et en rouge les fermes saines. Les deux premiers ratios

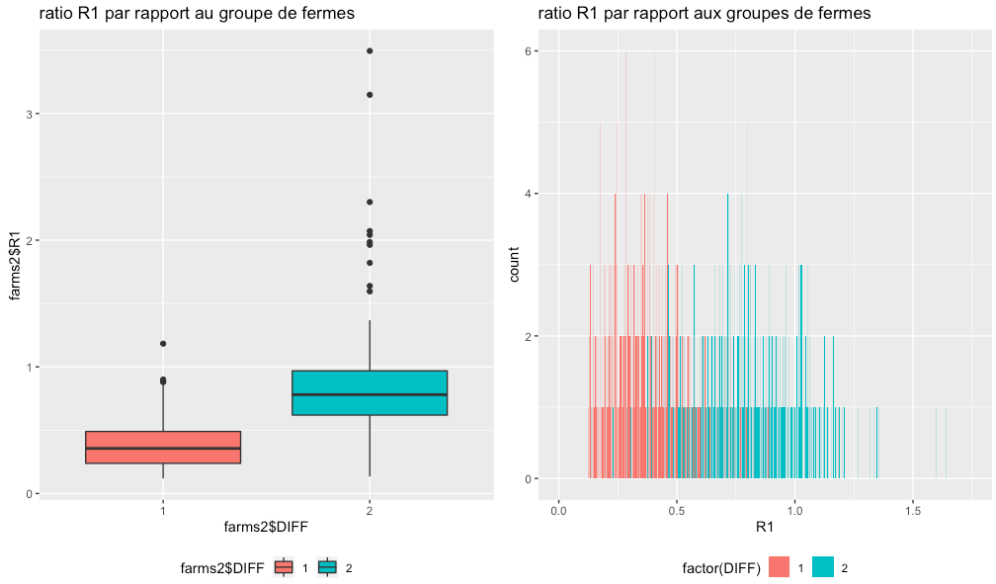


FIGURE 4 – Analyse du ratio R1

Ce ratio R1 quantifie la dette totale d'une ferme à la totalité des actifs, ainsi on remarque que plus ce ratio est important plus la ferme a tendance à être défaillante. En effet plus une ferme a de dettes plus son risque de faire faillite doit être important.

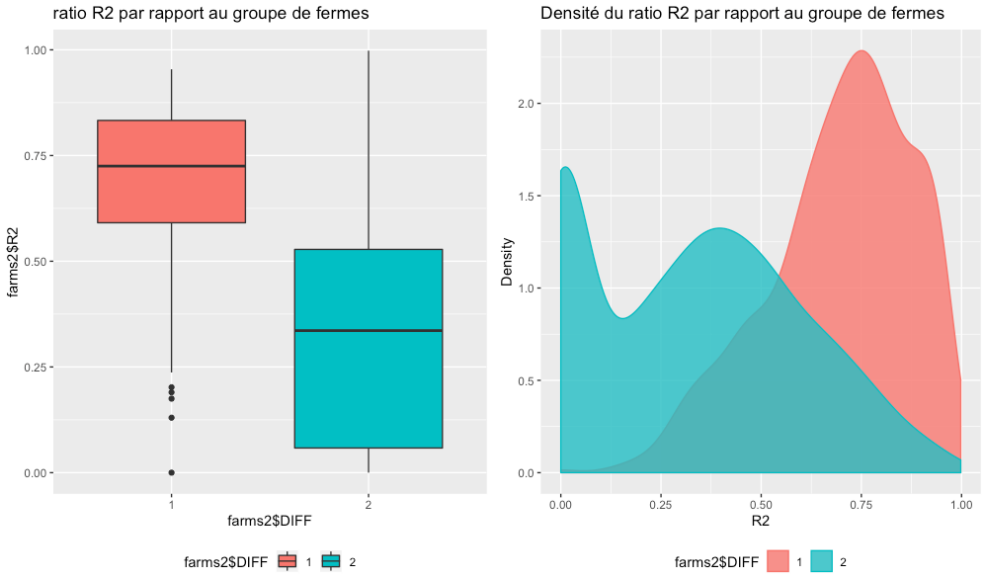


FIGURE 5 – Ratio R2

Ce ratio est appelé le ratio d'indépendance financière et met en avant l'état d'endettement financier de l'entreprise par rapport à ses fonds propres.

Ainsi comme nous pouvons le remarquer sur ces graphiques, quand ce ratio est faible la ferme aurait plus tendance à connaître des difficultés financières.

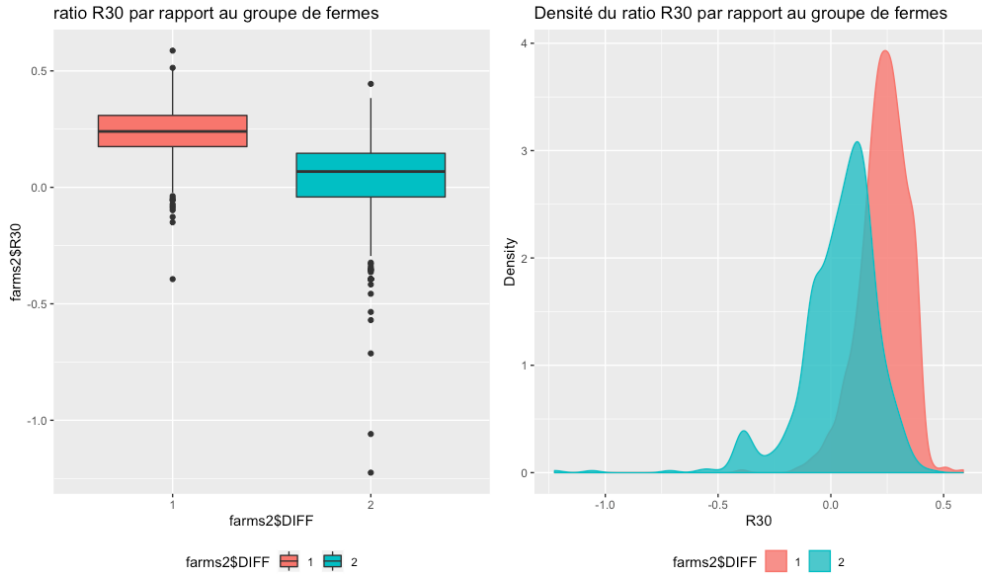


FIGURE 6 – Analyse du ratio R30

Le ratio quantifie les revenus de la ferme par rapport au produit brut.
D'après les graphiques, il semblerait que plus la valeur de ce ratio est grande moins la ferme à de risque d'appartenir au groupe défaillant et donc de faire faillite.

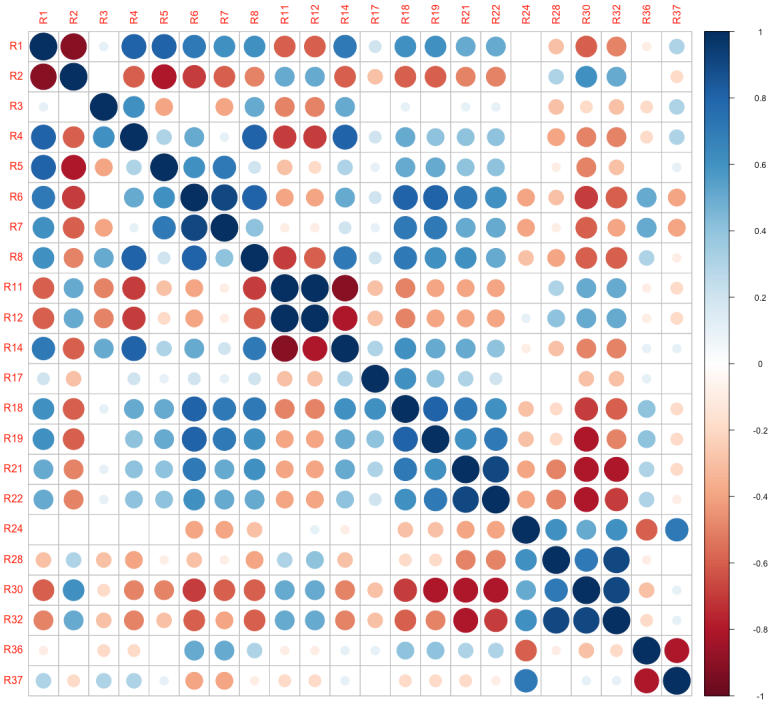
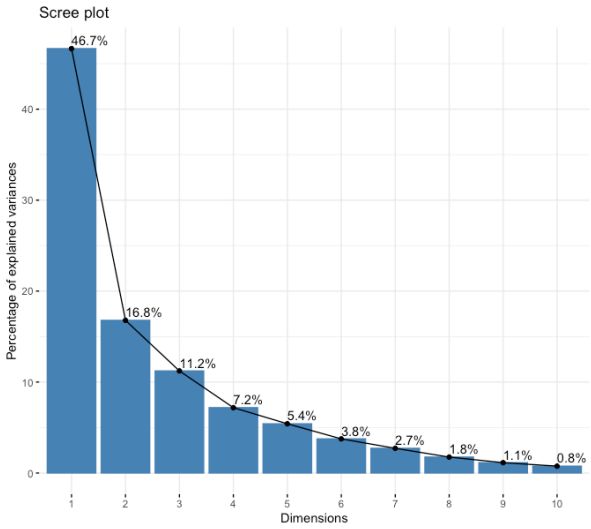


FIGURE 7 – Corrélation entre ratios financiers

Étude des corrélations Cette figure nous permet de remarquer les différentes corrélations entre ratios financiers on remarque une corrélation (positive ou négative) entre les ratios d'un même groupe mais également des corrélations assez fortes entre ratios de groupes différents. Certains ratios apportent donc peut-être des infirmations redondantes.
C'est ce que nous allons pouvoir voir grâce à l'analyse en composantes principales.

Analyse en composantes principales Pour résumer l'information contenue dans ces ratios financiers on fait une Analyse en composante principale. Ainsi on décide tout d'abord du nombre de composantes retenues.



	eigenvalue	percentage of variance	cumulative
comp 1	10.264935316	46.65879689	
comp 2	3.692949774	16.78613534	
comp 3	2.469801899	11.22637227	
comp 4	1.581180290	7.18718314	
comp 5	1.191289030	5.41495014	
comp 6	0.826545659	3.75702572	
comp 7	0.598010331	2.71822878	
comp 8	0.386044466	1.75474757	
comp 9	0.250359520	1.13799782	
comp 10	0.167045846	0.75929930	
comp 11	0.137037934	0.62289970	
comp 12	0.127529037	0.57967744	
comp 13	0.093675957	0.42579981	
comp 14	0.058505383	0.26593356	
comp 15	0.044486217	0.20221008	
comp 16	0.040318956	0.18326798	
comp 17	0.022568058	0.10258208	
comp 18	0.014627385	0.06648811	
comp 19	0.012013792	0.05460815	
comp 20	0.010473109	0.04760504	
comp 21	0.006359966	0.02890893	
comp 22	0.004242073	0.01928215	

FIGURE 8 – Histogramme des variances expliquées et Valeurs propres et pourcentage de variance expliqué

Selon la règle de Kaiser on devrait garder les 5 axes de rang supérieur (règle de la valeur propre supérieure à 1).
Cependant les axes F3, F4 et F5 sont assez délicats à analyser car assez spécifiques à certaines fermes. En effet même si les 5 axes rassemblent plus de 87 % de l'information totale nous allons garder le premier plan factoriel (les deux premiers axes) qui regroupent 63 % de l'inertie totale. Ce choix est confirmé par la règle du coude. (visible sur l'histogramme des variances expliquées).

Ainsi nous analysons uniquement les résultats de l' analyse en composantes principales sur le premier plan factoriel.

Il est possible de lire la corrélation des différents ratio grâce au cercle des corrélations ci-dessous.

Nous avons également choisi de représenter en variables supplémentaires les autres variables du jeu de données.

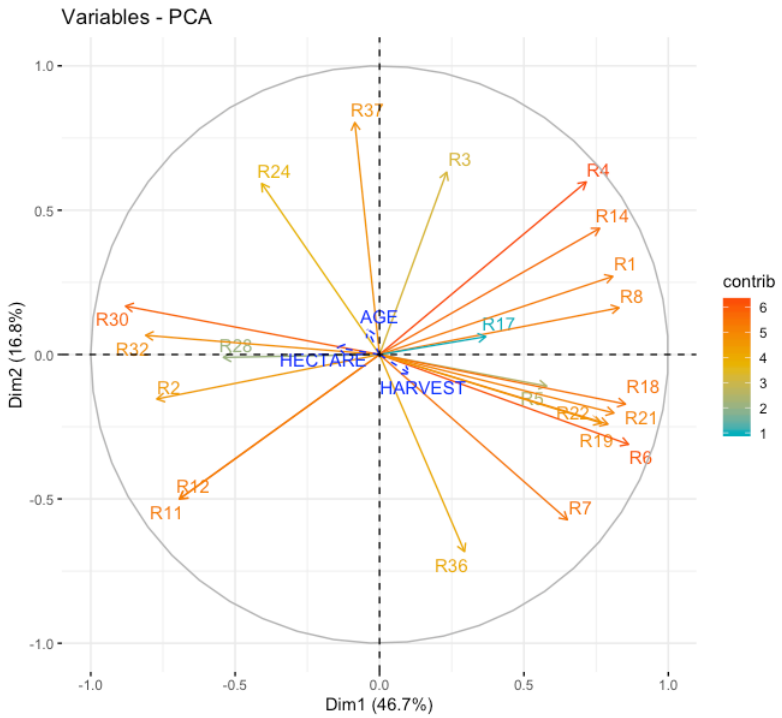


FIGURE 9 – ACP sur les deux premiers axes

Les variables quantitatives supplémentaires comme AGE, HECTARE sont trop peu corrélées avec les deux premiers facteurs. Ainsi leur projection sur le premier plan factoriel ne peut pas être interprétée.

L'axe F1 représente 47% de l'information totale. Cet axe montre l'opposition entre deux groupes de ratios :

Positivement corrélé	Négativement corrélé
R8 & R6	R30
R18 & R21	R32
R14 & R1	R2
R22	R11 & R12

— F1 $>$ 0 : regroupe des ratios mesurant les poids des dettes et des frais financiers d'une ferme,

— F1 $<$ 0 : regroupe des ratios mesurant la richesse d'une ferme : par leurs revenus avant les taxes, et leurs bénéfices

Ainsi les fermes ayant des problèmes financiers se trouvent sur l'axe F1$>$0.

L'axe F2 représente 17% de l'information totale, et montre l'opposition entre deux groupes de ratios :

Positivement corrélé	Négativement corrélé
R37	R36
R3	R7
R4	R11
R24	R12

Ainsi ce deuxième axe F2 représente la relation entre une efficacité productive plus ou moins importante et un montant plus ou moins élevé de dette courant par rapport à la dette totale.

Nous avons ci dessus la représentation des différentes fermes sur les plans F1-F2 :

FIGURE 10 – Représentation des fermes sur les plans F1-F2

On peut ainsi voir que les fermes saines sont bien séparées des fermes défaillantes sur ces axes. On remarque qu'une méthode de classification serait de créer une droite de regression : Nous obtenons l'équation de droite suivante :

$$y = -1.07578x + 0.00005$$

Les fermes au dessus de cette droite seront classées comme étant défaillantes alors que les fermes en dessous de la droite seront classées comme saines. C'est ce qui est réalisé par l'algorithme ci dessous :

```
DIFF_prevu_dim1 = rep(0,n)
for (i in 1:n){
  x = acp_dim1[["ind"]][["coord"]][i]
  point_pivot = 0.08202495
  if (x>point_pivot){  #si on est en dessous de la droite
    DIFF_prevu_dim1[i] <- 2
  }
  if (x<point_pivot){  #on est au dessus de la droite
    DIFF_prevu_dim1[i] <- 1
  }
}

obs_dim1 = DIFF_prevu_dim1
pred = acp_dim1[["call"]][["quali.sup"]][["quali.sup"]][["DIFF"]]
```

Nous obtenons ainsi deux vecteurs : le premier rendant les résultats de la classification des fermes par rapport à la droite. Le second représente les vrais valeurs et les vrais groupes pour chaque ferme. A partir de ces deux vecteur nous construisons une matrice de confusion avec les résultats ci dessous :

	$DIFF = 0$	$DIFF = 1$
$Y_{pred} = 0$	595	125
$Y_{pred} = 1$	58	482

— 91.11792 % des fermes saines sont bien classées

— 79.40692 % des fermes défaillantes sont bien classées

Nous représentons ci dessous les différentes fermes sur les plans F1-F3 et les plans F2-F3 pour remarquer un quelconque critère de distinction des deux groupes de fermes

FIGURE 11 – Représentation des fermes sur les plans F1-F3 (à gauche) et F2-F3 (à droite)

— Les fermes sont encore plutôt bien distinctes sur les plans F1-F3 (à gauche)

— Il est impossible de faire une distinction des deux groupes de fermes sur les places F2-F3 (à droite)

De ces graphiques, on en conclue que c'est l'axe 1 qui est indispensable à une bonne distinction des deux groupes de fermes. On crée une règle de classification utilisant seulement cet axe.

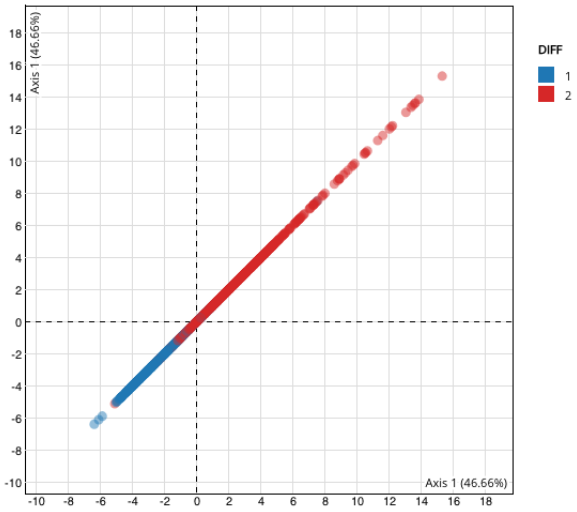


FIGURE 12 – Représentations des fermes sur le premier axe factoriel

La moyenne des individus du groupe des fermes saines pour l'axe F1 est : $\mu_0 = -2.16474485$. La moyenne des individus du groupe des fermes défaillantes pour l'axe F1 est $\mu_1 = 2.32879471$. Ainsi un point pivot est la moyenne de ces deux valeurs.

$$p = \frac{\mu_0 + \mu_1}{2} = 0.08202495$$

Ainsi, si la coordonnée d'une ferme sur l'axe F1 est supérieure à ce point pivot, nous l'attribuons au groupe des fermes défaillantes, au contraire si elle est inférieure à ce point pivot nous l'attribuons au groupe des fermes saines.

```
DIFF_prevu_dim1 = rep(0,n)
for (i in 1:n){
```

```
  x = acp_dim1[["ind"]][["coord"]][i]
  point_pivot = 0.08202495
  if (x>point_pivot){
    DIFF_prevu_dim1[i] <- 2
  }
  if (x<point_pivot){
    DIFF_prevu_dim1[i] <- 1
  }
}

obs_dim1 = DIFF_prevu_dim1
pred = acp_dim1[["call"]][["quali.sup"]][["quali.sup"]][["DIFF"]]
```

Nous obtenons à partir de cet algorithme le vecteur `obs_dim1` indiquant par des 1 : les fermes considérés comme saines et des 2 les fermes considérées comme défaillantes (par le point pivot).
Nous comparons ce vecteur prédit à l'actuel vecteur `DIFF` à l'aide d'un algorithme créant la matrice de confusion associée.

	$DIFF = 0$	$DIFF = 1$
$Y_{pred} = 0$	599	124
$Y_{pred} = 1$	54	483

Ainsi on obtient un taux de fermes saines bien classées de 91.7 %, et un taux de fermes défaillantes bien classées de 79.6 %.

L'utilisation de l'axe F1 donne alors de meilleurs résultats que l'utilisation des axes F1 et F2. Cependant avec cette méthode le taux de faux négatifs soit "observer une ferme saine alors qu'elle est défaillante" est trop élevé. Si on adopte le point de vue d'un banquier, ce modèle ne sera donc pas satisfaisant. En effet on veut éviter les situations de non remboursement des prêts, on préfère ainsi avoir un meilleur taux de classement pour les fermes défaillantes.

Cette ACP montre que la première composante principale s'avère être un facteur discriminant acceptable quand on cherche à privilégier l'intérêt général.

On peut cependant utiliser une méthode conçue pour maximiser la note globale des exploitations classées correctement pour améliorer les performances de classement.

3 Application de l'AFD

La méthode de construction de score l'analyse discriminante probabiliste Pour traiter l'exemple `farms` avec l'analyse factorielle discriminante on utilise seulement les ratios comme variables explicatives ; Nous avons jugé que les variables qui étaient supplémentaires en ACP n'étaient pas assez pertinentes pour la suite.

3.1 Choix des modèles et des variables

Modèle 1 : modèle complet

Modèle 2 : modèle utilisant le lambda de Wilks 0.01 .

```
greedy.wilks(DIFF~., data=farms, niveau=0.01)
```

```
Formula containing included variables:

DIFF ~ R1 + R32 + R14 + R17 + R2 + R3 + R36 + R21

Values calculated in each step of the selection procedure:

  vars Wilks.lambda F.statistics.overall
1   R1    0.5804430             909.3100
2  R32    0.5017071             624.2229
3  R14    0.4667025             478.4073
4  R17    0.4534737             378.1314
5   R2    0.4451038             312.6641
6   R3    0.4371952             268.8328
7  R36    0.4292739             237.7932
8  R21    0.4234460             212.9165
```

FIGURE 13 – Lambda de Wilks = 0.01

Modèle 3 : modèle utilisant le lambda de Wilks (0.05) .

```
greedy.wilks(DIFF~., data=farms, niveau=0.05)
```

```
Formula containing included variables:

DIFF ~ R1 + R32 + R14 + R17 + R2 + R3 + R36 + R21 + R7 + R18 +
    R19

Values calculated in each step of the selection procedure:

   vars Wilks.lambda F.statistics.overall
1    R1    0.5804430             909.3100
2   R32    0.5017071             624.2229
3   R14    0.4667025             478.4073
4   R17    0.4534737             378.1314
5    R2    0.4451038             312.6641
6    R3    0.4371952             268.8328
7   R36    0.4292739             237.7932
8   R21    0.4234460             212.9165
9    R7    0.4217347             190.4387
10  R18    0.4188155             173.3220
11  R19    0.4172071             158.4837
```

FIGURE 14 – Lambda de Wilks = 0.05

Autres modèles D'autres modèles sont ajoutés et correspondent à ceux trouvés par différents critères en regression logistique.

3.2 Analyse disriminante linéaire (LDA)

Nous implémentons l'algorithme ci dessous reprenant les modèles décrits précédemment, et rendant une matrice indiquant quels modèles répondent le plus aux critères. Le principe de cet algorithme sera ré-utilisé

pour l'analyse discriminante quadratique et la régression logistique.

```
N = 1000
scores_A = rep(0,8)

for (k in 1:N){
  sample = sample.split(DIFF, SplitRatio = 0.8)
  train = subset(farms, sample == TRUE)
  test = subset(farms, sample == FALSE)

  modele_1 = lda(DIFF~.,data=train)
  modele_2 = lda(DIFF ~ R1 + R3 + R14 + R17 + R36, data = train)
  modele_3 = lda(DIFF ~ R1 + R3 + R17 + R36, data = train)
  modele_4 = lda(DIFF ~ R1 + R14 + R17 + R36, data = train)
  modele_5 = lda(DIFF ~ R1 + R12 + R14 + R17 + R32 + R36,data=train)
  modele_6 = lda(DIFF ~ R2 + R7 + R17 + R32, data=train)
  modele_7 = lda(DIFF ~ R1 + R2 + R3 + R7 + R14 + R17 + R18 + R19 + R21 + R32 + R
     36,data=train)
  modele_8 = lda(DIFF ~ R1 + R2 + R3 + R14 + R17 + R21 + R32 + R36,data=train)

  liste_modeles = list(modele_1, modele_2, modele_3, modele_4,modele_5,modele_6,
     modele_7,modele_8)
  n = length(liste_modeles)

  A = matrix(0, nrow = 2, ncol=n)
  A[1,]= 1:n

  for (i in 1:n){
    diff.pred = predict(liste_modeles[[i]],test[,-1],method="predictive")
    erreur_pred = prop.table(table(diff.pred, test$DIFF))[2]   #taux de FN
    A[2,i] = erreur_pred
  }
  A_tri = A[,order(A[2,], decreasing = FALSE)]
  for (i in 1:n){
    scores_A[A_tri[1,i]] = scores_A[A_tri[1,i]] + (9-i)}
}

scores_A
#attribuer des scores: matrice deux lignes, une avec les modeles une avec les
    scores
```

L'algorithme ci-dessus partitionne le jeu de données en un échantillon d'apprentissage et un échantillon test à chaque passage dans la boucle.

A chaque répétition, plusieurs modèles sont créés et leurs capacités prédictives sont notées grâce à un score calculé en fonction de la capacité du modèle à minimiser le taux de faux négatif.

Tous les scores obtenus sont additionnés pour chaque modèle et le vecteur 'Scores A' contient donc à la fin de l'algorithme la somme des scores obtenus pour tous les modèles. On retient alors celui ayant le score le plus élévé.

Dans notre exemple, nous retenons donc le modèle 5 comme modèle optimal. Soit le modèle contenant les ratios `R1`, `R12`, `R14`, `R17`, `R32` et `R36`.

On veut pouvoir afficher les histogrammes de score pour ce modèle pour pouvoir déterminer le seuil à choisir donnant un taux d'erreur minimum à celui fixé par l'analyste pour ce genre d'étude. Nous construisons donc 2 dataframes contenant les scores des fermes saines et des fermes défaillantes.

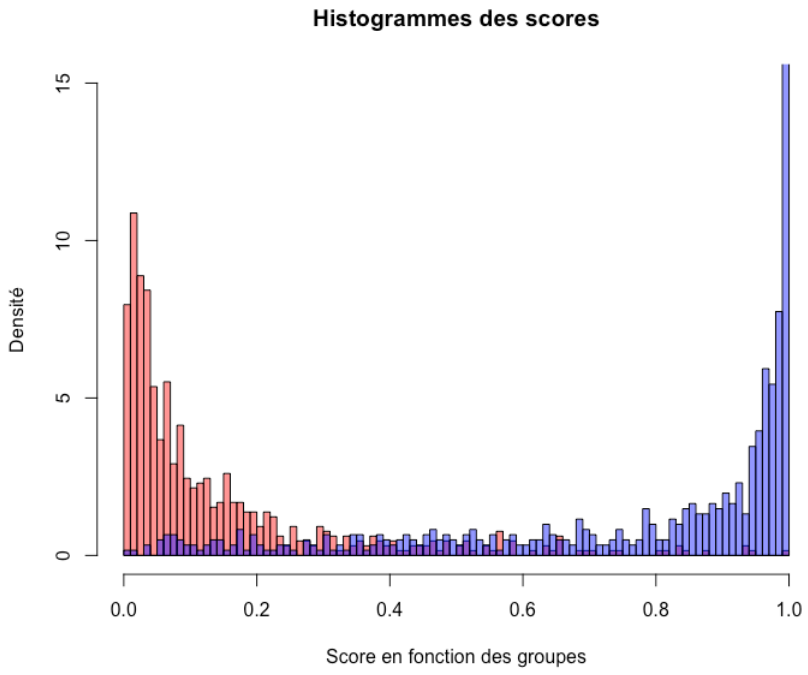


FIGURE 15 – Histogrammes des scores de la méthode LDA

Une fois les histogrammes affichés, on se rend bien compte qu'un seuil de 0.5 comme il est coutume de prendre de manière générale donnerait beaucoup d'erreurs de classement. Nous cherchons donc grâce aux commandes suivantes le seuil à choisir qui donne un taux d'erreur fixé par l'analyste.

```
ApproxQuantile(hgB2, 0.05) #rend le seuil qu'il faut consid rer pour n'avoir que
    5% de risque de se tromper lors de l'affectation a un des deux groupes
ApproxQuantile(hgB2, 0.1) # idem mais pour 10%
```

Pour n'avoir que 5% de risque de mal classer une ferme défaillante, il faudrait choisir un seuil de 0.148. Pour n'avoir que 10% de risque de mal classer une ferme défaillante, il faudrait choisir un seuil de 0.305.

Ainsi on recompile l'algorithme avec un de ces deux seuils à la place de 0.5.

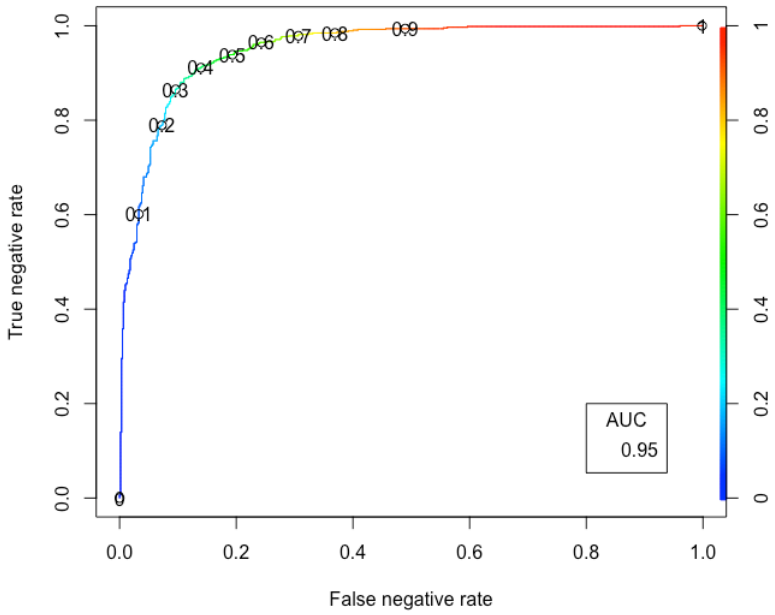


FIGURE 16 – Histogrammes des scores de la méthode LDA

Nous obtenons une belle courbe ROC quasi parfaite, le modèle obtenu est très satisfaisant pour classer les fermes. De plus l'aire sous la courbe est très proche de 1 ce qui confirme la qualité de notre modèle optimal.

3.3 Analyse discriminante quadratique (QDA)

Nous compilons la même procédure que pour l'analyse discriminante linéaire avec les mêmes modèles et obtenons un nouveau modèle optimal : le modèle complet.

Cependant les histogrammes de score et les seuils obtenus sont inexploitables, ainsi nous préférons comparer le modèle optimal trouvé par lda au même modèle mais dans sa version quadratique.

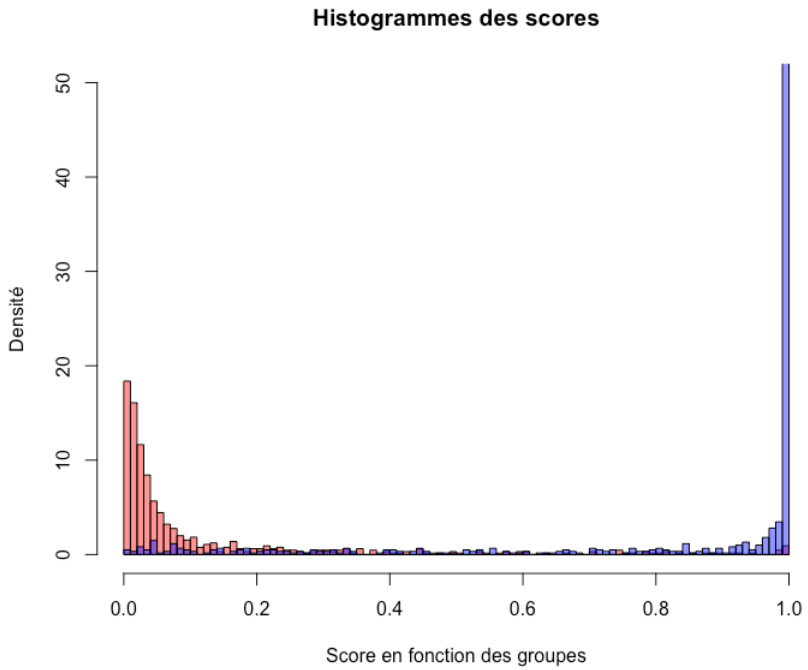


FIGURE 17 – Histogrammes des scores de la méthode LDA

Les histogrammes de score sont ainsi plus lisibles.

Avec la version lda, pour n'avoir que 5% de risque de mal classer une ferme défaillante, on trouvait un seuil de 0.148, ici ce seuil est de 0.075. Le deux valeurs sont donc très différentes et il est difficilement envisageable de prendre un seuil aussi bas.

Pour n'avoir que 10% de risque de mal classer une ferme défaillante, on trouvait un seuil de 0.305, là où ce seuil vaut 0.21 maintenant. Ce seuil est déjà plus envisageable. Ainsi nous recompilons l'algorithme avec ce nouveau seuil, et obtenons toujours le modèle final. On peut donc tracer la courbe ROC pour vérifier la qualité du modèle

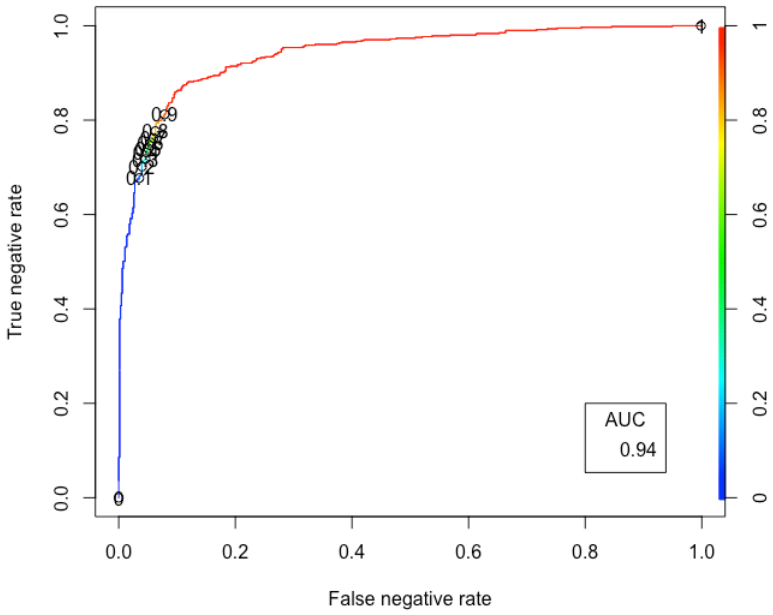


FIGURE 18 – Histogrammes des scores de la méthode QDA

L'aire sous la courbe (AUC) est proche de 1 (0.94) et la courbe est proche du coin supérieur gauche donc c'est un bon modèle. Nous traitons ce même exemple avec les mêmes modèles dans le cas de la régression logistique.

Troisième partie

Régression logistique

1 Introduction

Les racines de la régression logistique plongent loin dans l'histoire de l'analyse des données. En effet c'est vers 1840 que P.F Verhulst introduit ce qu'il appelle "équation logistique" pour répondre à une problématique de dynamique des populations . La régression logistique consiste à expliquer une variable Y qualitative (variable cible), par une ou plusieurs variables explicatives X_j(qualitatives ou quantitatives). Cette méthode a été introduite en 1944 par Berkson [1] en biostatistiques.
La méthode de régression logistique est très appréciée pour sa généralité, son interprétabilité et sa robustesse.
La fonction logistique est utilisée dans de nombreux domaines :

- épidémiologie : diffusion d'une épidémie
- marketing : ventes d'un nouveau produit
- psychologie : pour prédire certains comportements
- technologie

Nous nous concentrons dans ce rapport au cas où la variable cible Y est binaire. On suppose qu'il y a donc deux groupes à discriminer. Ainsi la variable à expliquer Y prend deux modalités 0 ou 1.
Quand le nombre de modalités de la variable à expliquer est supérieur à 2 on parle de régression logistique *polytomique* (scrutin a plus de deux candidats, degrés de satisfaction pour un produit, mention a un examen....)

Dans plusieurs cas la régression logistique sera privilégiée par rapport à l'analyse factorielle discriminante car elle nécessite moins d'hypothèses. En effet l' avantage de cette méthode est qu'il n'y a pas besoin d'hypothèses de multinormalité.

Notations On note :

- Y la variable à expliquer est à valeurs dans $\{0 \quad 1\}$.
- $X = \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix} \in \mathbb{R}^{p+1}$ un vecteur de variables explicatives $X_j \quad \forall j \in [\![1, p]\!]$
- Le vecteur des coefficients $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}$ à estimer par maximum de vraisemblance.

1. Joseph BERKSON (1899,1982) était un physicien, médecin statisticien américain. Il a introduit la notion de régression logistique dans son article *Are there two regressions ? (1950)*

— $x = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \in \mathbb{R}^{p+1}$ une réalisation de X, y est une réalisation de Y, y_i suit une loi de Bernoulli de paramètre $\pi_\beta(x_i)$.

— $(X_1, Y_1), \cdots (X_n, Y_n)$ est un n-échantillon aléatoire et de même loi que le couple (X, Y)

— $(x_1, y_1), \cdots, (x_n, y_n)$ une réalisation de $(X_1, Y_1) \cdots (X_n, Y_n)$

L'objectif de la régression logistique est de modéliser l'espérance conditionnelle de Y par rapport à X : $\mathbb{E}[Y \mid X = x]$.

En régression linéaire, on a :

$$\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Ceci ne convient pas lorsque Y est binaire (0 ou 1) puisque le terme ci dessus est non borné alors que $\mathbb{P}(Y \mid X = x)$ est dans l'intervalle $[0, 1]$. On a alors quand Y est binaire (0 ou 1) :

$$\mathbb{E}[Y \mid X = x] = 1 \times \mathbb{P}(Y = 1 \mid X = x) + 0 \times \mathbb{P}(Y = 0 \mid X = x)$$

Ainsi :

$$\begin{aligned} \mathbb{E}[Y \mid X = x] &= \mathbb{P}(Y = 1 \mid X = x) \\ &= f_\beta(x) \end{aligned}$$

Cette expression est la probabilité a posteriori d'appartenir au premier groupe. La fonction $f_\beta(x)$ est appelée *fonction de transfert*.

Exemple avec un cas simple (une variable explicative) On veux essayer d'expliquer la présence de maladie cardiovasculaire par une seule variable explicative : l'âge du patient.

Ici on va donc expliquer la variable `CHD` (0 si le patient est sain, 1 sinon) par la variable `AGE`. On dispose de 100 individus.

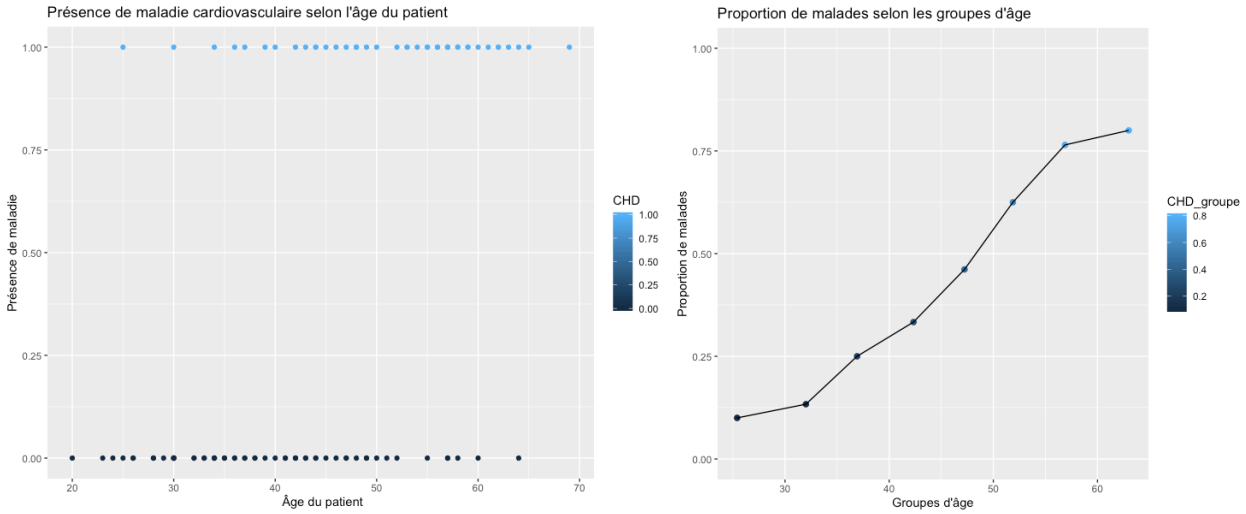


FIGURE 19 – Représentation des individus à gauche, proportion de malades selon des groupes d'âge à droite

On aperçoit grâce au graphique de gauche qu'il est difficile de tirer des conclusions quant à l'influence de l'âge sur l'apparition de maladie cardiovasculaire à cause de la variabilité de la variable CHD. Ainsi nous regroupons les individus par classes d'âge prédéfinies.

On remarque que la liaison entre les deux variables est plus claire sur le second graphique (à droite) grâce à cette répartition par classes d'âges. En effet plus l'âge augmente plus le risque de contracter une maladie cardiovasculaire est élevé. On remarque par ailleurs que la forme suit une courbe sigmoïde en forme de "S".

C'est pour cela qu'on va utiliser une fonction sigmoïde. Dans la régression logistique cette fonction est $\pi_\beta(x)$:

$$\pi_\beta(x) = \mathbb{P}(G1|x) = \mathbb{P}(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

La fonction $\pi_\beta(x)$ est comprise dans $]0, 1[$, elle convient donc à une probabilité et donne souvent une bonne représentation des phénomènes.

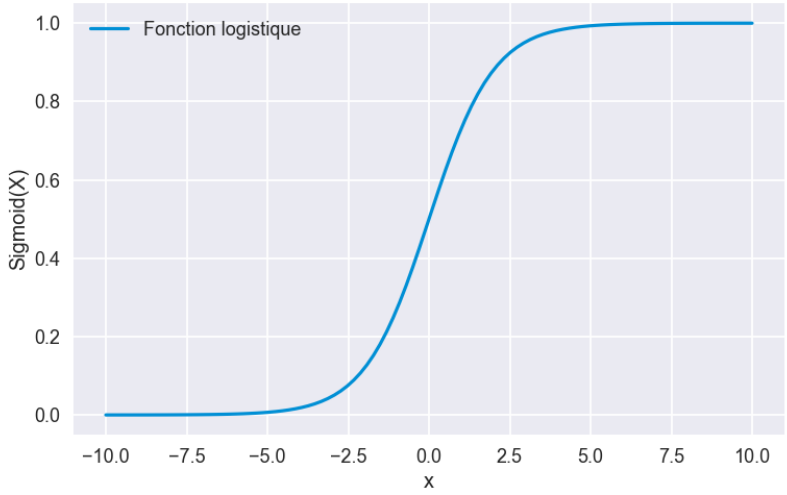


FIGURE 20 – Représentation graphique de la fonction logistique sur Python - Sigmoïde

Définition 1.1

La fonction $\pi_\beta(x)$ est appelée *fonction logistique*. Sa représentation graphique est une sigmoïde en fonction des valeurs de x.

$$\begin{aligned} \pi_\beta(x) : \quad \mathbb{R}^{p+1} &\longrightarrow]0, 1[\\ x &\longmapsto \pi_\beta(x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \end{aligned}$$

On cherche à écrire l'espérance conditionnelle de la variable à expliquer Y comme combinaison linéaire de variables explicatives X.

Définition 1.2

Soit Y une variable à valeurs dans $\{0,1\}$ à expliquer par p variables explicatives $X_1, \cdots, X_p$.
Le modèle logistique propose une modélisation de la loi de $Y \mid X = x$ par une loi de Bernoulli de paramètre $\pi_\beta(x) = \mathbb{P}_\beta(Y = 1 \mid X = x)$ telle que :

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

ou encore

$$\text{logit}(\pi_\beta(x)) = x^T\beta$$

La fonction logit est appelée *fonction de lien*, elle est bijective, dérivable sur $]0,1[$ à valeurs dans $\mathbb{R}$.

Utilisée avec la fonction de logarithme néperien, logit est la réciproque de $f(x) = \dfrac{1}{1+e^{-x}}$ qui est utilisée pour linéariser les fonctions logistiques.

1.1 Interprétation avec des Odds-ratio

L'odds ratio permet de mesurer l'effet d'un facteur. L'odds ratio d'une variable explicative quantifie l'évolution d'un phénomène lorsque X_j passe de x à $x+1$ toutes variables étant égales par ailleurs.

$$Odds = \frac{\pi(x)}{1-\pi(x)}$$

On définit l'Odds ratio comme :

$$OR = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = e^{\beta_i}$$

— Si $\beta_j < 0 \iff OR < 1$ cela indique que la variable explicative a une influence négative sur la variable à prédire.
— Si $\beta_j > 0 \iff OR > 1$ cela indique que la variable explicative a une influence positive sur la variable à prédire.
— Si $\beta_j = 0 \iff OR = 1$ alors la variable explicative n'a aucune influence sur la variable à prédire

Quand la variable explicative X_j est binaire, on a :

$$Odds = \frac{\mathbb{P}(Y = 1 \mid X_j = 1)}{\mathbb{P}(Y = 0 \mid X_j = 1)}$$

On obtient un seul odds ratio qui est :

$$OR = \frac{\dfrac{\mathbb{P}(Y = 1 \mid X_j = 1)}{\mathbb{P}(Y = 0 \mid X_j = 1)}}{\dfrac{\mathbb{P}(Y = 1 \mid X_j = 0)}{\mathbb{P}(Y = 0 \mid X_j = 0)}} = e^{\beta_j}$$

C'est le facteur par lequel on multiplie la côte lorsque x passe de 0 à 1.

Retour sur l'exemple introductif .

```
cardio.glm = glm(CHD~AGE,family=binomial)
summary(cardio.glm)
```

```
Call:
glm(formula = CHD ~ AGE, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
AGE          0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35

Number of Fisher Scoring iterations: 4
```

FIGURE 21 – Sortie R de la fonction glm

On peut lire les coefficients $\widehat{\beta_0} = -5.30945$ et $\widehat{\beta_1} = 0.11092$. On va utiliser ces coefficients pour tracer la sigmoïde représentant la proportion de malades selon les âges :

On obtient la fonction logistique suivante :

$$\pi_\beta(x) = \frac{e^{-5.31+0.11x}}{1+e^{-5.31+0.11x}}$$

On peut la linéariser de la manière suivante pour obtenir la fonction logit de lien représentée sur la figure ci dessous à droite :

$$\text{logit}(\pi_\beta(x)) = -5.31 + 0.11x$$

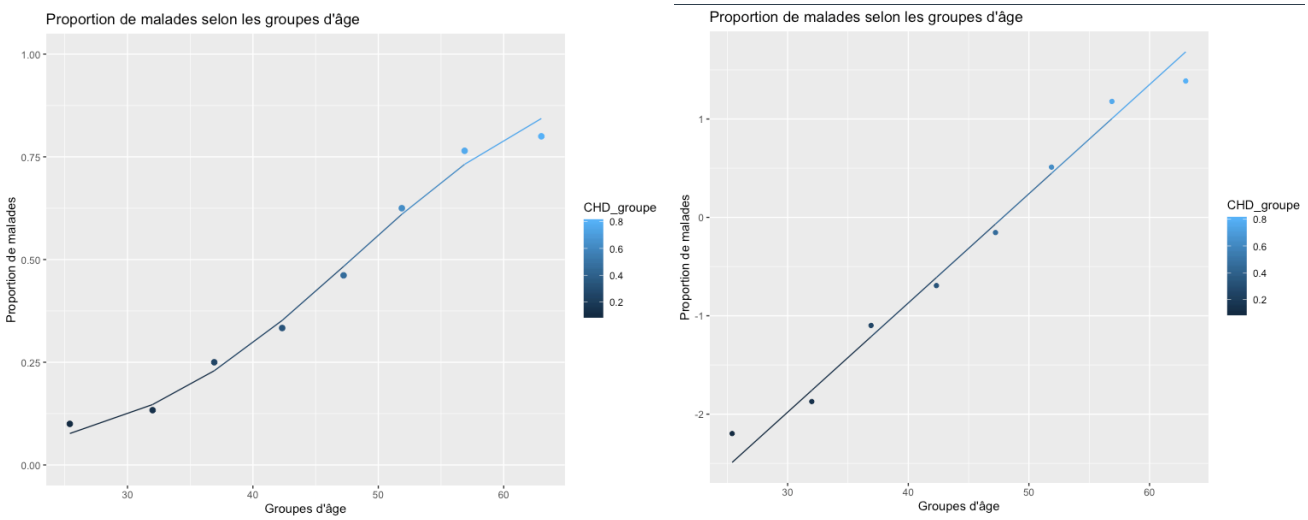


FIGURE 22 – Fonction logistique et fonction logit des malades selon les groupes d'âge

Les coefficients $\widehat{\beta}_j$ s'interprètent comme des logarithmes népériens d'odds ratio. Ainsi on a :

$$OR = e^{\hat{\beta}_1} = e^{0.1109} \approx 1.12$$

Donc quand un patient vieillit d'un an, son risque de contracter une maladie cardiovasculaire est multiplié par 1.12.

2 Estimation des paramètres

Pour estimer le vecteur de paramètres β on utilise la méthode de maximum de vraisemblance à partir d'un échantillon *iid* de n observations. En effet la variable Y à expliquer étant qualitative, on ne peut pas utiliser la méthode d'estimation par les moindres carrés habituelle.

La vraisemblance La vraisemblance pour une observation (y_i, x_i) peut s'écrire :

$$\ell(\beta; y_i, x_i) = \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

Comme les observations sont *iid* on peut écrire que la vraisemblance du n-échantillon est égale au produit des vraisemblances par observation :

$$\ell(\beta; y, x) = \prod_{i=1}^{p} \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

La log vraisemblance

Proposition 2.1

La log vraisemblance s'écrit

$$\beta \longmapsto \ell\ell_X(\beta; y, x) = \sum_{i=1}^{n} y_i \ln\left(\pi_\beta(x_i)\right) + (1 - y_i) \ln\left(1 - \pi_\beta(x_i)\right)$$

Démonstration. La vraisemblance s'écrit :

$$\ell(\beta; y, x) = \prod_{i=1}^{n} \pi_\beta(x_i)^{y_i} \times (1 - \pi_\beta(x_i))^{1-y_i}$$

Or $\pi_\beta(x_i) \in]0, 1[$. Donc la vraisemblance est strictement positive, on peut calculer la log vraisemblance.

$$\begin{aligned}
\ell\ell(\beta) = \ln \ell(\beta) &= \sum_{i=1}^{n} \ln\left(\mathbb{P}\left(Y = y_i \mid X = x_i\right)\right) \\
&= \sum_{i=1}^{n} y_i \ln\left(\pi_\beta(x_i)\right) + (1 - y_i) \ln\left(1 - \pi_\beta(x_i)\right)
\end{aligned}$$

□

Équations de vraisemblance Le vecteur gradient au point β est défini par :

$$\nabla_\beta \ell\ell(\beta) = \begin{pmatrix} \frac{\partial \ell\ell}{\partial \beta_0}(\beta) \\ \vdots \\ \frac{\partial \ell\ell}{\partial \beta_p}(\beta) \end{pmatrix}$$

Calculons $\dfrac{\partial \ell\ell}{\partial \beta_j}(\beta) \quad \forall j \in [\![0, p]\!]$. On a :

$$\begin{aligned}
\ell\ell(\beta) &= \sum_{i=1}^{n} y_i \ln\left(\pi_\beta(x_i)\right) + (1 - y_i) \ln\left(1 - \pi_\beta(x_i)\right) \\
&= \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}\right) + (1 - y_i) \ln\left(1 - \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}\right) \\
&= \sum_{i=1}^{n} y_i \ln\left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}\right) + (1 - y_i) \ln\left(\frac{1}{1 + e^{\beta^T x_i}}\right)
\end{aligned}$$

Avec $\beta^T x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}$ et $x_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix}$

$$\begin{aligned}
\frac{\partial}{\partial \beta_j}\left[\ln\left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right)\right] &= \frac{x_{ij}e^{\beta x_i}\left(1+e^{\beta^T x_i}\right) - e^{\beta^T x_i}\left(x_{ij}e^{\beta^T x_i}\right)}{\left(1+e^{\beta^T x_i}\right)^2} \times \frac{1+e^{\beta^T x_i}}{e^{\beta^T x_i}} \\
&= \frac{x_{ij}e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2} \times \frac{1+e^{\beta^T x_i}}{e^{\beta^T x_i}} \\
&= \frac{x_{ij}}{1+e^{\beta^T x_i}} \\
\frac{\partial}{\partial \beta_j}\left[\ln\left(\frac{1}{1+e^{\beta^T x_i}}\right)\right] &= -\frac{x_{ij}e^{\beta^T x_i}}{\left(1+e^{\beta^T x_i}\right)^2} \times \left(1+e^{\beta^T x_i}\right) \\
&= -\frac{x_{ij}e^{\beta^T x_i}}{1+e^{\beta^T x_i}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell\ell}{\partial \beta_j}(\beta) &= \sum_{i=1}^{n} y_i \frac{x_{ij}}{1+e^{\beta^T x_i}} - (1-y_i)\frac{x_{ij}e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \\
&= \sum_{i=1}^{n} \frac{x_{ij}\left(y_i - e^{\beta^T x_i} + y_i e^{\beta^T x_i}\right)}{1+e^{\beta^T x_i}} \\
&= \sum_{i=1}^{n} x_{ij}\frac{y_i\left(1+e^{\beta^T x_i}\right) - e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \\
&= \sum_{i=1}^{n} x_{ij}\left(y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}\right) \\
&= \sum_{i=1}^{n} x_{ij}\left(y_i - \pi_\beta(x_i)\right) \qquad \forall j \in [\![0,p]\!] \text{ avec } x_{i0}=1
\end{aligned}$$

On obtient l'écriture générale :

$$\nabla_\beta \ell\ell(\beta) = \sum_{i=1}^{n} x_i(y_i - \pi_\beta(x_i)) = \begin{pmatrix} \sum_{i=1}^{n}(y-\pi_\beta(x_i)) \\ \sum_{i=1}^{n} x_{i1}(y-\pi_\beta(x_i)) \\ \vdots \\ \sum_{i=1}^{n} x_{ip}(y-\pi_\beta(x_i)) \end{pmatrix}$$

On peut également l'écrire sous forme matricielle :

$$X^T(Y - \Pi_\beta)$$

avec $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}$ $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ et $\Pi_\beta = \begin{pmatrix} \pi_\beta(x_1) \\ \vdots \\ \pi_\beta(x_n) \end{pmatrix} \in \mathbb{R}^n$

Recherche d'estimateur du maximum de vraisemblance :
Si l'estimateur de maximum de vraisemblance $\widehat{\beta}$ existe, il est solution de l'équation :

$$X^T (Y - \Pi_\beta) = 0$$

Ainsi rechercher les solutions de cette équation revient à résoudre $p+1$ équations à $p+1$ inconnues $(\beta_0, \beta_1, \cdots, \beta_p)$:

$$\begin{cases} y_1 + \cdots + y_n = \pi_\beta(x_1) + \cdots + \pi_\beta(x_n) & j = 0 \\ x_{1j}y_1 + \cdots + x_{nj}y_n = x_{1j}\pi_\beta(x_1) + \cdots + x_{nj}\pi_\beta(x_n), & \forall j \in [\![1, p]\!] \end{cases}$$

$$\iff \begin{cases} y_1 + \cdots + y_n = \dfrac{e^{\beta^T x_1}}{1 + e^{\beta^T x_1}} + \cdots + \dfrac{e^{\beta^T x_n}}{1 + e^{\beta^T x_n}} & j = 0 \\ x_{1j}y_1 + \cdots + x_{nj}y_n = x_{1j}\dfrac{e^{\beta^T x_1}}{1 + e^{\beta^T x_1}} + \cdots x_{nj}\dfrac{e^{\beta^T x_n}}{1 + e^{\beta^T x_n}}, & \forall j \in [\![1, p]\!] \end{cases}$$

Ce système d'équations n'a pas de solution analytique et se résout par des procédures de calcul numérique (Newton Raphson, algorithme IRLS, ...).

Théorème 2.1

Si X est de rang maximal, la log vraisemblance $\beta \longmapsto \ell\ell(\beta)$ est strictement concave : $\widehat{\beta}$ existe et est unique.

Démonstration. Calculons la matrice Hessienne de la log vraisemblance

$$\nabla^2_\beta \ell\ell(\beta; y, x) = \left(\frac{\partial^2 \ell\ell}{\partial\beta_i \partial\beta_j}(\beta; y, x) \right)_{1\leq i,j\leq p}$$

$$\begin{aligned}
\nabla^2_\beta \ell\ell(\beta; y, x) &= \nabla_\beta \left(\nabla_\beta \ell\ell(\beta; y, x)\right) \\
&= \nabla_\beta \left(\sum_{i=1}^{n} x_i (y_i - \pi_\beta(x_i)) \right) \\
&= -\sum_{i=1}^{n} x_i \nabla_\beta(\pi_\beta(x_i)) \\
&= -\sum_{i=1}^{n} x_i x_i^T \frac{e^{\beta^T x_i}}{\left(1 + e^{\beta^T x_i}\right)^2} \\
&= -\sum_{i=1}^{n} x_i x_i^T \pi_\beta(x_i)(1 - \pi_\beta(x_i))
\end{aligned}$$

Or $\pi_\beta(x_i)(1 - \pi_\beta(x_i)) > 0$ car $\pi_\beta(x_i) \in]0, 1[$.

De plus $x_i^T x_i = \|x_i\|^2$ donc $\|x_i\|^2 \geq 0$ et $\|x_i\|^2 = 0$ pour $x_i = 0$.

Sous forme matricielle on peut écrire, en posant $\pi_i = \pi_\beta(x_i)$:

$$\begin{aligned}
H(\beta; Y, X) = \nabla^2_\beta \ell\ell(\beta; Y, X) &= -\begin{pmatrix}
\sum_{i=1}^{n} \pi_i(1-\pi_i) & \sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) \\
\sum_{i=1}^{n} x_{i1}\pi_i(1-\pi_i) & \sum_{i=1}^{n} (x_{i1})^2\pi_i(1-\pi_i) & \cdots & \sum_{i=1}^{n} x_{i1}x_{ip}\pi_i(1-\pi_i) \\
\vdots & & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ip}\pi_i(1-\pi_i) & \cdots & \cdots & \sum_{i=1}^{n} (x_{ip})^2\pi_i(1-\pi_i)
\end{pmatrix} \\
&= -\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}^T \begin{pmatrix} \pi_1(1-\pi_1) & & 0 \\ & \ddots & \\ 0 & & \pi_n(1-\pi_n) \end{pmatrix} \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \\
&= -X^T \Delta_\beta X
\end{aligned}$$

Δ_β est une matrice $n \times n$ diagonale où le k-ième terme est égal à $\pi_k(1 - \pi_k) > 0$.
De plus si X est de rang maximal ($rg(X) = p + 1$) alors X est injective et la matrice Δ_β est définie positive.

Ainsi la matrice hessienne de la log vraisemblance est définie négative, alors la log vraisemblance est strictement concave par rapport à β. Ceci garantit, l'unicité du maximum de cette fonction. Ainsi quel que soit le choix des conditions initiales ou de l'algorithme utilisé, les estimateurs du maximum de vraisemblance convergeront vers la vraie valeur $\widehat{\beta}$. □

L'algorithme le plus souvent utilisé pour le calcul de cet estimateur est l'algorithme de Newton Raphson. Enfin puisqu'on utilise l'estimateur du maximum de vraisemblance, il est possible de construire des intervalles

de confiance asymptotiques utiles pour les tests de sélection de variables.

La théorie du maximum de vraisemblance nous donne la loi asymptotique des estimateurs : il est donc possible de tester la significativité des variables explicatives : Trois tests sont généralement utilisés (voir protocole) :

— Le test de Wald
— Le test du rapport des vraisemblances ou déviance
— le test du score

Ces deux premiers tests ont été mentionnés dans le protocole.

3 Applications de la régression logistique

3.1 Fonction GLM

Pour la régression logistique, on utilise avec le logiciel R la fonction `glm` : modèle linéaire généralisé. Cette fonction possède la syntaxe suivante :

```
glm(formula, family=familytype(link=linkfunction), data=dataset)
```

L'approche de la fonction GLM consiste à :

— choisir une loi pour $Y \mid X = x$ en fonction du nombre de modalités de Y.
— choisir une fonction de lien $g_\beta(x)$ bijective et dérivable.
— réaliser une transformation de l'espérance conditionnelle $\mathbb{E}[Y \mid X = x]$ par la fonction g_β pour obtenir notre fonction en sigmoïde :

$$g_\beta\left(\mathbb{E}[Y \mid X = x]\right) = f_\beta(x) = x^T\beta$$

Pour estimer les coefficients $\beta_0, \cdots \beta_p$ la fonction `glm` utilise l'algorithme de Newton Raphson.

Choix	Logistique	Log-linéaire	Linéaire
Loi de $Y \mid X = x$	Bernoulli	Poisson	Normale
Modélisation de $\mathbb{E}[Y \mid X = x]$	$\text{logit}\,\mathbb{E}[Y \mid X = x] = x^T\beta$	$\log \mathbb{E}[Y \mid X = x] = x^T\beta$	$\mathrm{E}[\ Y \mid X = x] = x^T\beta$

3.2 Exemple économique

Pour second modèle d'application nous continuons l'exemple sur les exploitations fermières et le classement de celles ci dans deux groupes : le groupe des fermes saines et celui des fermes défaillantes, à l'aide de plusieurs ratios financiers.

3.2.1 Choix des modèles

Modèle 1 : Modèle complet .

```
modele_complet = glm(DIFF ~., data = farms_ordre, family = binomial(link = "logit
    "))
summary(modele_complet)
```

```
Call:
glm(formula = DIFF ~ ., family = binomial(link = "logit"), data = farms_ordre)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5135  -0.3737  -0.0845   0.2435   3.0521

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6932     4.3190  -2.707  0.00678 **
R1           24.5373     8.3017   2.956  0.00312 **
R2            2.3209     3.5047   0.662  0.50783
R3            4.3735     1.8713   2.337  0.01943 *
R4          -10.1236     7.7739  -1.302  0.19283
R5          -13.4190     8.4799  -1.582  0.11355
R6           -1.1654     1.6788  -0.694  0.48754
R7            2.0817     1.9851   1.049  0.29433
R8           -2.6752     2.0528  -1.303  0.19251
R11          -1.0817     2.3500  -0.460  0.64529
R12           1.5486     1.6158   0.958  0.33786
R14           2.3297     0.7781   2.994  0.00275 **
R17          39.0878     9.9563   3.926 8.64e-05 ***
R18         -10.0287    17.5203  -0.572  0.56705
R19           2.3088    10.8064   0.214  0.83082
R21           1.1110     3.9905   0.278  0.78070
R22           1.2595     1.8532   0.680  0.49674
R24           2.2843     5.6556   0.404  0.68629
R28         -12.1249    11.6916  -1.037  0.29971
R30           7.8600    11.1415   0.705  0.48052
R32          -1.8340    13.9176  -0.132  0.89516
R36           1.4477     0.4830   2.998  0.00272 **
R37          -2.1438     1.7772  -1.206  0.22772
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1745.1  on 1259  degrees of freedom
Residual deviance:  658.5  on 1237  degrees of freedom
AIC: 704.5

Number of Fisher Scoring iterations: 8
```

FIGURE 23 – Étude statistique du modèle complet

On peut à partir de ces résultats et des variables les plus significatives sélectionner un nouveau modèle contenant les ratios : `R1`, `R3`, `R14`, `R17` et `R36`

Modèle 2 : Modèle contenant les variables les plus significatives .

Ce modèle comprend les variables `R1`, `R3`, `R14`, `R17` et `R36`

Nous obtenons le même modèle avec le test de la déviance et le test de Wald.

Modèle 3 : Modèle minimisant l'AIC .

Le modèle minimisant l'AIC contient 11 ratios.

```
Call:  glm(formula = DIFF ~ R1 + R3 + R8 + R12 + R14 + R17 + R22 + R28 +
    R30 + R36 + R37, family = binomial(link = "logit"), data = farms_ordre)

Coefficients:
(Intercept)           R1           R3           R8          R12          R14          R17          R22
    -9.9854      10.2464       5.3125      -3.2304       0.9784       2.5304      34.7873       1.8935
        R28          R30          R36          R37
   -12.0852       8.0662       1.4516      -1.4783

Degrees of Freedom: 1259 Total (i.e. Null);  1248 Residual
Null Deviance:     1745
Residual Deviance: 663.7  AIC: 687.7
```

FIGURE 24 – Fonction step rendant le modèle minimisant l'AIC

Autres modèles On rajoute les modèles trouvés lors de l'analyse discriminante soit ceux utilisant le critère du lambda de Wilks.

3.2.2 Recherche du modèle optimal

Le modèle optimal réalisant le meilleur score est le modèle minimisant l'AIC ainsi contenant 11 ratios.

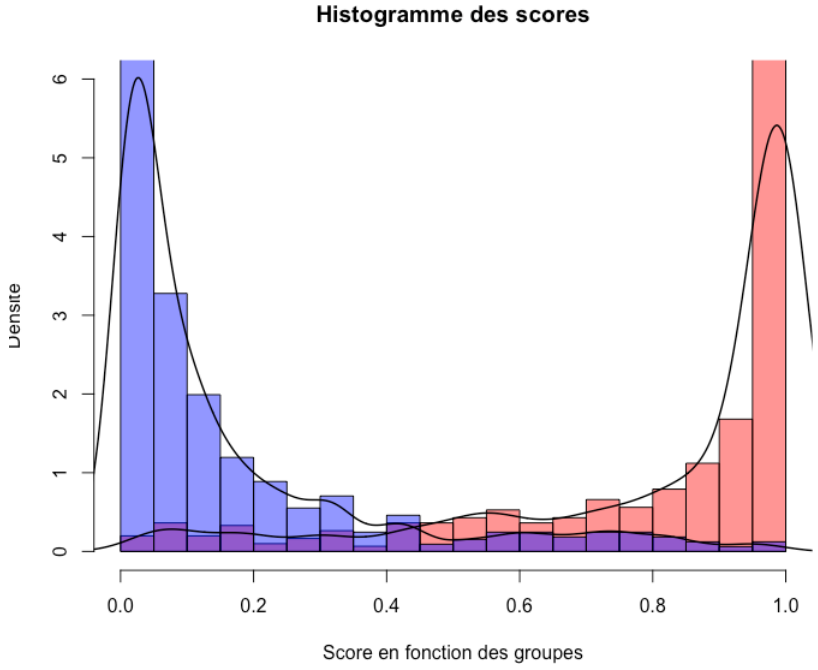


FIGURE 25 – Histogramme des scores pour le modèle optimal minimisant l'AIC

On obtient la courbe ROC suivante :

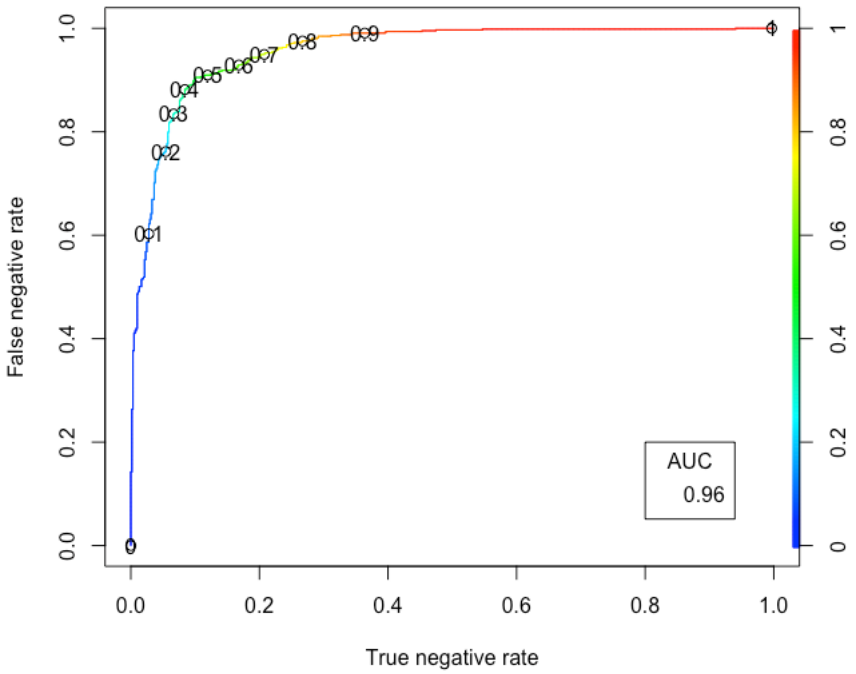


FIGURE 26 – Courbe ROC et AUC pour le modèle optimal

La courbe ROC pour ce modèle optimal donnent des résultats très satisfaisants. En effet l'aire sous la courbe est très proche de 1.

A ce stade, nous pouvons comparer les performances de chaque méthodes de scoring que nous avons effectuées ici. Les valeurs des AUC sont très proches pour chaque méthode (lda,qda et regression logistique). La meilleure semble être la régression logistique pour cet exemple. De plus, nous avons dû supposer les données gaussiennes et les matrices de variance-covariance égale lors de l'analyse discriminante linéaire alors que ce n'est pas nécessaire pour la régression logistique, ce qui représente un avantage.

Cependant, l'analyse discriminante linéaire rend un modèle optimal se constituant de 6 ratios contre 11 en régression logistique. Donc si un AUC de 0.95 est considéré comme acceptable, la lda sera donc préférée car nécessitant moins de ratio à connaitre pour classer une ferme dans le bon groupe.

3.3 Exemple médical : le diabète

La base de données `diabete` contient des observations de 768 individus. Tous les patients ici sont des femmes d'au moins 21 ans d'origine indienne Pima en Amérique du Nord. Le peuple Pima est connu pour être une des communautés comportant le plus grand pourcentage d'obèses et de diabétiques au monde, et à ce titre est un sujet d'études pour les scientifiques.

Introduction Avant de commencer, faisons un rapide point sur le métabolisme du glucose et l'apparition de diabète. Le corps utilise le glucose comme carburant pour certains de ses organes. Ce sucre est apporté par l'alimentation. L'insuline est une hormone qui sert à réguler ce taux de glucose dans le sang, elle est utilisée en cas d'hyperglycémie. Elle permet au glucose d'entrer dans les cellules du corps où il sera utilisé sous forme

d'énergie ou mis en réserve pour être libéré quand le corps en aura besoin (en cas d'hypoglycémie). Chez les personnes saines l'insuline est produite de façon continue selon les besoins.

Un dérèglement de ce métabolisme est la cause d'une maladie : le diabète.
Deux types de diabète existent :

- de type I : dû a une absence de production d'insuline et concerne 10 % des diabétiques.
- de type II : l'insuline est produite mais n'est pas reconnue par les différents récepteurs des cellules, ainsi son taux diminue dans le temps. Il concerne 90% des diabétiques.

Enfin plusieurs facteurs environnementaux ou génétiques sont responsables de cette maladie. Ainsi le but de notre étude est de déterminer quels facteurs décrivent et prédisent le mieux l'apparition de diabète. Cela peut être utile pour savoir quels tests peuvent être utilisés pour la détection de cette maladie.

L'objectif est de prédire si la patiente est diabétique ou non.
Ce qui se traduit par expliquer la variable Y (`Outcome`) par les variables explicatives quantitatives suivantes :

- X_1 = `Pregnancies` : Nombre de grossesses de la patiente. *Un diabète particulier existe et est lié à la grossesse : le diabète gestationnel , il peut apparaitre pendant une grossesse et les femmes l'ayant contracter deviennent à risque de développer un diabète de type 2 dans les années qui suivent. Ainsi plus une femme a eu de grossesses plus son risque de développer des diabètes gestationnels augmente*
- X_2 = `Glucose` : Concentration de glucose plasmatique après 2 heures par un test de tolérance au glucose par voie orale.
- X_3 = `BloodPressure` : Pression artérielle diastolique (mm Hg) *Les diabétiques auraient plus tendance à connaître des problèmes d'hypertension*
- X_4 = `SkinThickness` : Épaisseur du pli cutané au niveau du triceps (mm)
- X_5 = `Insulin` : mesure de l'insuline 2h après une injection d'insuline (mu U/ml)
- X_6 = `BMI` : Indice de masse corporelle : $\frac{\text{poids en kg}}{(\text{ taille en m })^2}$. *Le surpoids et l'obésité augmentent le risque de développer un diabète de type II. Alors que les individus atteints de diabète de type I sont des individus généralement maigres.*
- X_7 = `DiabetesPedigreeFunction` : score qui représente la probabilité d'être diabétique selon les antécédents familiaux. *Des facteurs génétiques peuvent engendrer un diabète.*
- X_8 = `Age` : âge de la patiente au moment du diagnostic. *Le diabète de type II a tendance à se manifester quand l'individu est âgé, alors que le diabète de type I concerne les individus plus jeunes.*

Analyse du jeu de données .

```
> head(diabete)
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

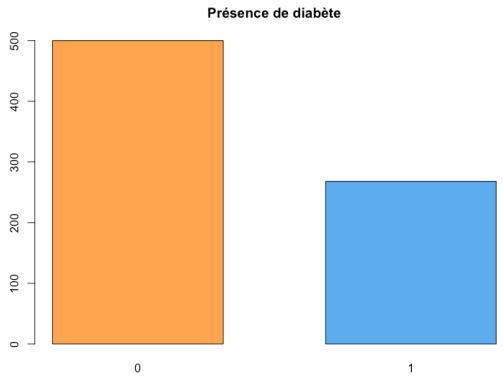


FIGURE 27 – Représentation des données des 6 premières patientes (à gauche) Présence de diabète ou non chez les patientes (à droite)

Dans cet échantillon 268 patientes sont diabétiques contre 500 non diabétiques.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Min. : 0.000	Min. : 0.0	Min. : 0.00	Min. : 0.00	Min. : 0.0	Min. : 0.00	Min. :0.0780	Min. :21.00	0:500
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 62.00	1st Qu.: 0.00	1st Qu.: 0.0	1st Qu.:27.30	1st Qu.:0.2437	1st Qu.:24.00	1:268
Median : 3.000	Median :117.0	Median : 72.00	Median :23.00	Median : 30.5	Median :32.00	Median :0.3725	Median :29.00	
Mean : 3.845	Mean :120.9	Mean : 69.11	Mean :20.54	Mean : 79.8	Mean :31.99	Mean :0.4719	Mean :33.24	
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:32.00	3rd Qu.:127.2	3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00	
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.0	Max. :67.10	Max. :2.4200	Max. :81.00	

FIGURE 28 – Descriptif des variables

On remarque que certaines valeurs de ces variables sont assez atypiques. En effet on peut remarquer que de nombreuses valeurs pour les variables `Skin Thickness` et `Insulin` (soit l'épaisseur de la peau et le taux d'insuline) sont nulles. C'est ce que l'on remarque en observant le minimum et premier quartile respectif de ces deux variables. Or il est impossible que l'épaisseur de la peau soit strictement nul. Il est également très étonnant de voir autant de valeurs nulles pour la variable `Insulin`. On s'est tout d'abord demandé si ces individus ayant un taux d'insuline nul n'étaient pas des diabétiques de type 1. Cependant après des recherches sur le peuple Pima on a remarqué que ce qui les caractérisait était leur prévalence du diabète de type 2 élevée mais aussi le taux d'obésité.
De plus la variable `Insulin` dans cette base de données correspond à de l'insuline qui a été injectée auparavant, et non de l'insuline créée par le corps même de l'individu. Il est donc difficilement envisageable de trouver un taux d'insuline nul 2h après injection.

On en conclut donc que ces valeurs nulles correspondent à des valeurs manquantes. Représentant plus de 30% du jeu de données, les enlever conduirait selon nous, à beaucoup d'information perdue. Nous allons donc les traiter.

Traitement des valeurs manquantes .

```
list(Column = colSums(diabete==0))
```

```
$Column
            Pregnancies                 Glucose           BloodPressure           SkinThickness                 Insulin
                    111                       5                      35                     227                     374
                    BMI DiabetesPedigreeFunction                     Age                 Outcome
                     11                       0                       0                     500
```

FIGURE 29 – Nombre de valeurs manquantes supposées par variable

En observant les autres valeurs de ce tableau nous avons remarqué que les variables `Glucose`, `BloodPressure` et `BMI` contenaient également des valeurs manquantes : en effet il est impossible d'avoir un IMC, un taux de glucose ou encore une pression artérielle diastolique nulle.

La variable `Pregnancies` contient elle aussi des valeurs nulles cependant nous ne pouvons pas avancer la nature de ces valeurs : représent-t-elles des femmes n'ayant connu aucune grossesse ou cette variable contient-elle, elle aussi des valeurs manquantes ? Il est complexe de répondre à cette question, ainsi nous décidons que ces valeurs nulles représentent l'absence de grossesse.
Grâce à ces remarques il faudra être vigilants quant aux résultats de nos analyses.

Ainsi pour traiter et estimer ces valeurs manquantes, nous utilisons une méthode des plus proches voisins c'est-à-dire l'algorithme KNN (*K nearest neighbors*) avec 5 groupes, et obtenons ainsi un nouveau jeu de données `diabete_new`.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35.00000	189.74880	33.6	0.627	50	1
2	1	85	66	29.00000	60.56782	26.6	0.351	31	0
3	8	183	64	30.64939	272.43145	23.3	0.672	32	1
4	1	89	66	23.00000	94.00000	28.1	0.167	21	0
5	0	137	40	35.00000	168.00000	43.1	2.288	33	1
6	5	116	74	21.28276	115.69694	25.6	0.201	30	0

FIGURE 30 – Représentation des données des 6 premières patientes après traitement des données manquantes

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Min. : 0.000	Min. : 44.0	Min. : 24.00	Min. : 7.00	Min. : 14.00	Min. :18.20	Min. :0.0780	Min. :21.00	0:500
1st Qu.: 1.000	1st Qu.: 99.0	1st Qu.: 64.00	1st Qu.:22.00	1st Qu.: 89.21	1st Qu.:27.50	1st Qu.:0.2437	1st Qu.:24.00	1:268
Median : 3.000	Median :117.0	Median : 72.00	Median :29.00	Median :130.00	Median :32.30	Median :0.3725	Median :29.00	
Mean : 3.845	Mean :121.7	Mean : 72.32	Mean :29.02	Mean :152.14	Mean :32.45	Mean :0.4719	Mean :33.24	
3rd Qu.: 6.000	3rd Qu.:140.2	3rd Qu.: 80.00	3rd Qu.:35.00	3rd Qu.:187.28	3rd Qu.:36.60	3rd Qu.:0.6262	3rd Qu.:41.00	
Max. :17.000	Max. :199.0	Max. :122.00	Max. :99.00	Max. :846.00	Max. :67.10	Max. :2.4200	Max. :81.00	

FIGURE 31 – Analyse exploratoire des différentes variables après traitement des données manquantes

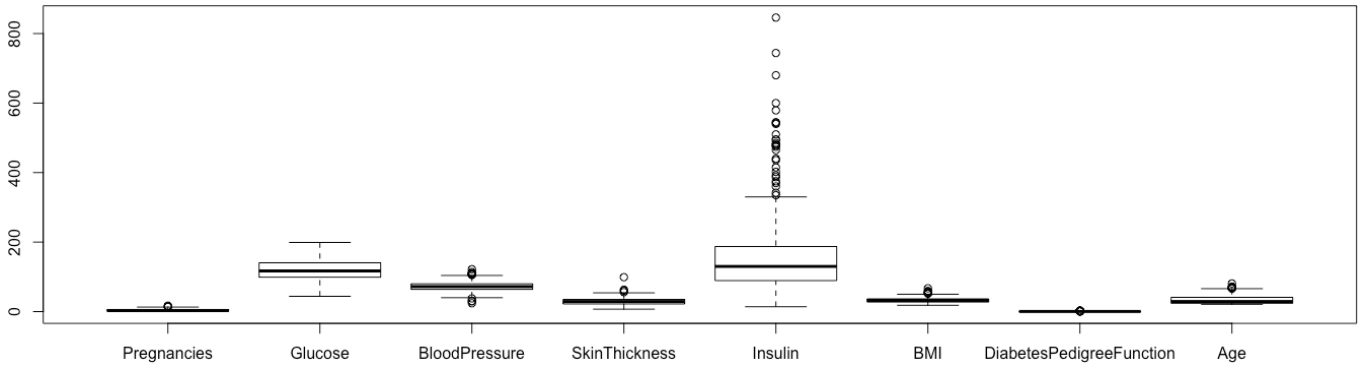


FIGURE 32 – Boites à moustache des variables explicatives

Analyse exploratoire du nouveau jeu de données En analysant ces box plot on aimerait transformer la variable Insuline en logarithme pour un des modèles. Cette variable contient beaucoup d'outliers et l'étendue prise par celle ci est plus grande que pour les autres variables.
De plus nous analysons les effets de quelques autres variables explicatives : `Age` et `Glucose`.

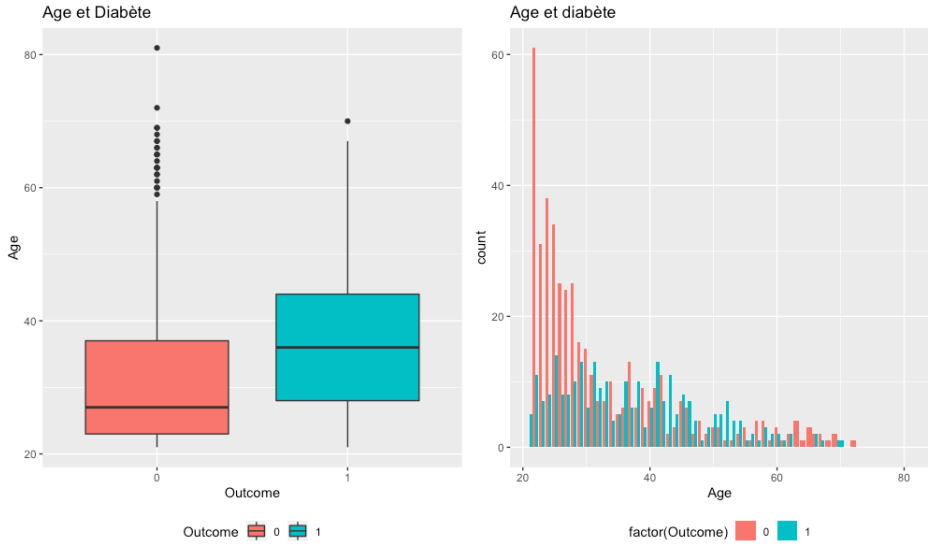


FIGURE 33 – Effet de l'âge sur l'apparition du diabète

On remarque que les individus diabétiques (en bleu) sont en moyenne plus âgés que les individus non diabétiques (en rouge).

Jusqu'à 35 ans il y a en moyenne plus d'individus non diabétiques que diabétiques.

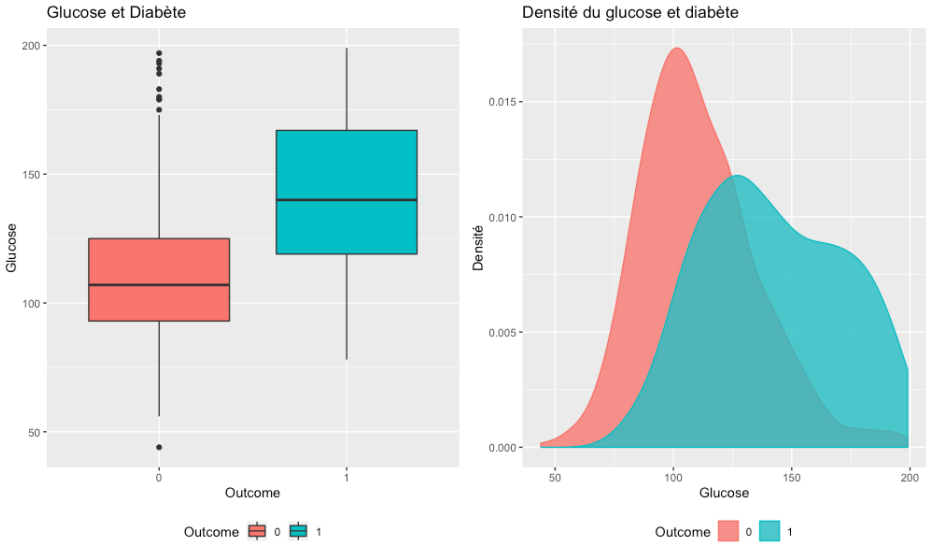


FIGURE 34 – Effet du glucose sur l'apparition de diabète

Le taux de glucose sanguin 2h après un repas est en moyenne plus élevé chez un individu diabétique. En effet, par définition même de cette maladie chronique l'individu reste longtemps en hyperglycémie.

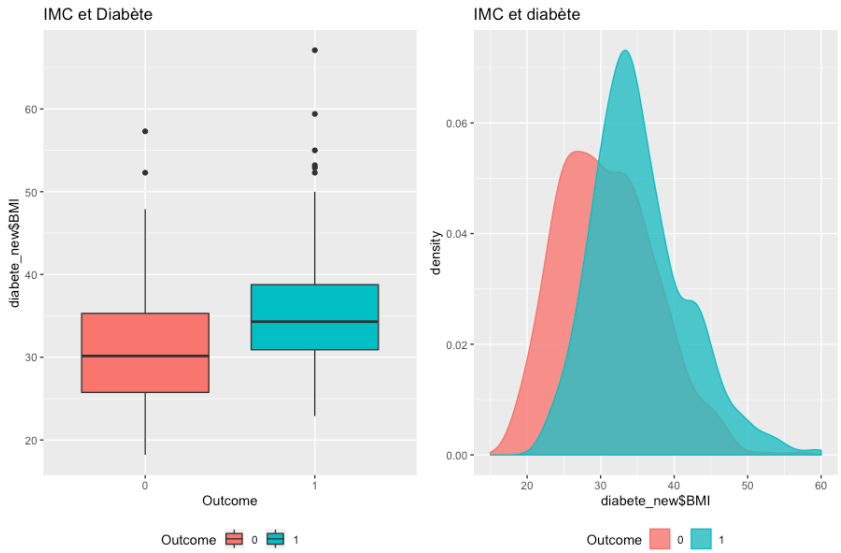


FIGURE 35 – Effet de l'IMC sur l'apparition de diabète

```r
surpoids = filter(diabete_new, BMI>25 & BMI<30)    #176 individus
obesite = filter(diabete_new, BMI>30)              #473 individus
```

On parle de surpoids quand l'indice de masse corporel est compris entre $25kg/m^2$ et $30kg/m^2$. On parle d'obésité quand celui ci dépasse 30 kg/m^2.
Ainsi 84 % des femmes Pimas dans cet échantillon ont un poids supérieur à la normale. Parmi ces femmes 73 % sont même obèses.
On remarque également qu'en moyenne les individus ayant un indice de masse corporel élevé sont diabétiques. Cela confirme également le fait que le diabète dont sont victimes les femmes Pima serait un diabète de type 2.

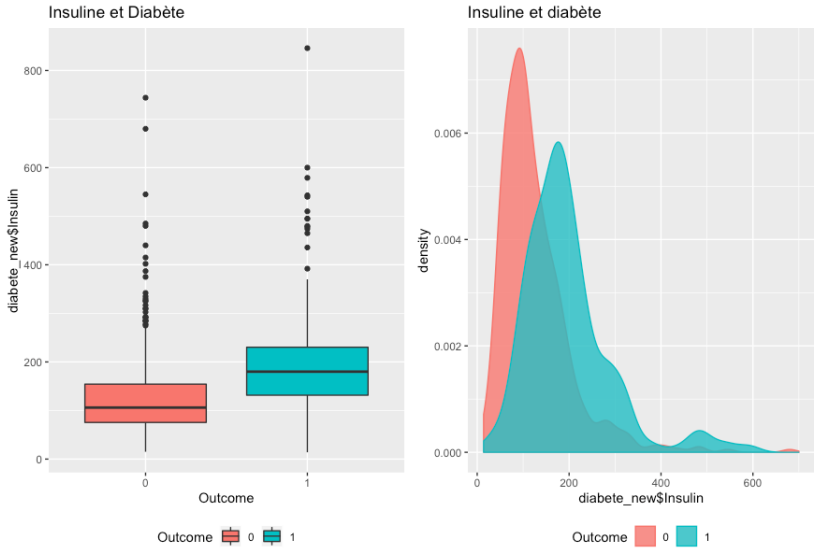


FIGURE 36 – Effet de l'insuline sur l'apparition de diabète

Le taux d'insuline après 2h d'injection est plus élevé chez des individus diabétiques que non diabétiques. Le diabète de type 2 est un diabète dit insulino-résistant, les cellules d'un individu souffrant de cette pathologie reconnaissent de moins en moins l'hormone hypoglycémiante par conséquent le glucose ne rentre donc pas dans les différents organes. Le glucose et l'insuline se trouvent donc en quantité plus élevée dans le sang chez une patiente malade.

Étude des corrélations Dans un but d'explorer nos données mais aussi réfléchir à de futurs modèles, nous menons une étude des corrélations entre les variables explicatives.

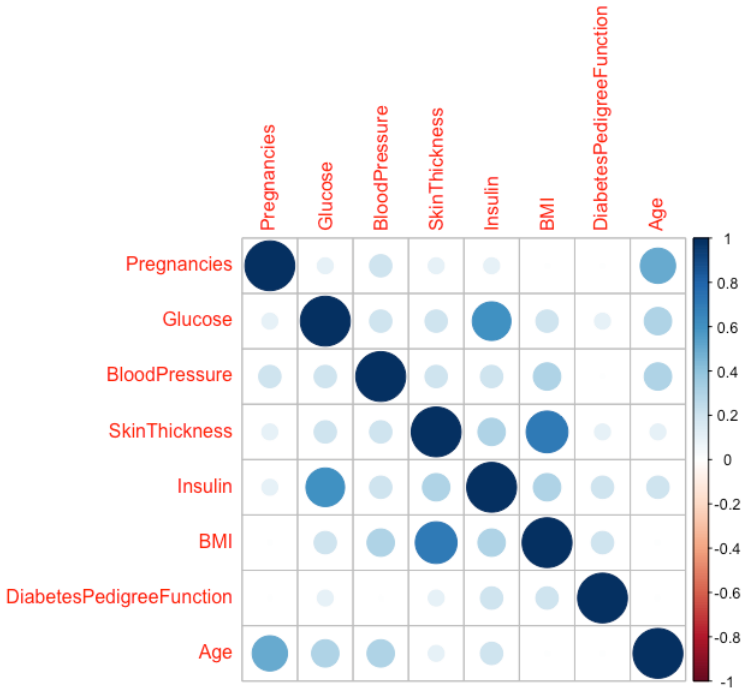


FIGURE 37 – Corrélation entre variables

— La variable `Age` est corrélée positivement à la variable `Pregnancies`. En effet plus une femme a d'enfants plus elle a tendance à être âgée.

— La variable `Insulin` est corrélée positivement à la variable `Glucose`. En effet d'une part quand un individu sain est en hyperglycémie (i.e le taux de glucose dans le sang est élevé) des cellules du pancréas libèrent dans le sang l'hormone hypoglycémiante : Insuline. Ainsi plus il y a de glucose dans le sang plus l'hormone d'insuline est libérée. D'autre part un individu souffrant de diabète de type 2 produit bien de l'insuline (quand il est jeune) mais ses cellules développent une résistance à celle ci ainsi il n'y a pas d'action hypoglycémiante. Quant aux diabétiques de type I

— La variable `BMI` est corrélée positivement à la variable `SkinThickness`. Plus un individu a un indice de masse corporel élevé plus sa peau a tendance à être épaisse.

3.3.1 Choix des modèles

Modèle 1 : Modèle complet Notre premier modèle comme pour chaque étude est notre modèle complet.

```
modele_complet = glm(Outcome ~ .  , data = diabete, family = binomial(link = "
    logit"))
summary(modele_complet)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7943  -0.7198  -0.3950   0.7099   2.3821

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.0755034  0.8330402 -10.894  < 2e-16 ***
Pregnancies               0.1239076  0.0324565   3.818 0.000135 ***
Glucose                   0.0366408  0.0042206   8.681  < 2e-16 ***
BloodPressure            -0.0082190  0.0085586  -0.960 0.336896
SkinThickness             0.0055866  0.0136126   0.410 0.681514
Insulin                  -0.0001204  0.0012111  -0.099 0.920814
BMI                       0.0892859  0.0201685   4.427 9.56e-06 ***
DiabetesPedigreeFunction  0.8690125  0.2982066   2.914 0.003567 **
Age                       0.0126303  0.0094929   1.331 0.183353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 713.27  on 759  degrees of freedom
AIC: 731.27

Number of Fisher Scoring iterations: 5
```

FIGURE 38 – Résultat de la fonction glm pour le modèle complet

Nous pouvons interpréter les coefficients comme des logarithmes d'odds ratio. Ainsi quand une femme du peuple Pima a une grossesse en plus son risque de contracter le diabète est multiplié par $e^{\widehat{\beta_1}} = e^{0124} = 1.13$.

Nous décidons à partir de ces résultats de créer un nouveau modèle : ce nouveau modèle comportera les variables les plus significatives au seuil de 5 % soit : `Pregnancies`, `Glucose`, `BMI` et `DiabetesPedigreeFunction`.

Modèle 2 : Variables significatives

$$\text{logit}(\pi_\beta(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_6 X_6 + \beta_7 X_7$$

Ce modèle ressort également comme le meilleur modèle lors de l'utilisation du test de Wald et de la déviance comme le montrent les sorties suivantes :

```
Anova(modele_complet, type = 3,test.statistic = "Wald")
Anova(modele_complet, type = 3, test.statistic = "LR")
```

```
Analysis of Deviance Table (Type III tests)

Response: diabete_new$Outcome
                        LR Chisq Df Pr(>Chisq)
Pregnancies               15.121  1  0.0001008 ***
Glucose                   90.239  1  < 2.2e-16 ***
BloodPressure              0.923  1  0.3367366
SkinThickness              0.169  1  0.6807448
Insulin                    0.010  1  0.9208769
BMI                       20.525  1  5.887e-06 ***
DiabetesPedigreeFunction   8.678  1  0.0032199 **
Age                        1.758  1  0.1848390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Deviance Table (Type III tests)

Response: diabete_new$Outcome
                          Df    Chisq Pr(>Chisq)
(Intercept)                1 118.6887  < 2.2e-16 ***
Pregnancies                1  14.5744  0.0001347 ***
Glucose                    1  75.3663  < 2.2e-16 ***
BloodPressure              1   0.9222  0.3368956
SkinThickness              1   0.1684  0.6815135
Insulin                    1   0.0099  0.9208142
BMI                        1  19.5982  9.556e-06 ***
DiabetesPedigreeFunction   1   8.4921  0.0035668 **
Age                        1   1.7702  0.1833526
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 39 – Modèle avec les variables utilisant le test de déviance (à gauche) et le test de Wald(à droite)

Le modèle des variables significatives vérifie également les critères de sélection automatique BIC et AIC.

```
bestglm(diabete, family = binomial, IC = "AIC")
```

```
AIC
BICq equivalent for q in (0.204884522523828, 0.93459410869902)
Best Model:
                           Estimate Std. Error    z value     Pr(>|z|)
(Intercept)             -9.18954044 0.70562425 -13.023278 9.021385e-39
Pregnancies              0.14271089 0.02753313   5.183244 2.180599e-07
Glucose                  0.03690915 0.00348911  10.578386 3.753703e-26
BMI                      0.08861285 0.01471174   6.023276 1.709217e-09
DiabetesPedigreeFunction 0.89287591 0.29527948   3.023833 2.495940e-03
```

FIGURE 40 – Modèle avec les variables minimisant le critère de l'AIC

```
bestglm(diabete, family = binomial, IC = "BIC")
```

```
BIC
BICq equivalent for q in (0.204884522523828, 0.93459410869902)
Best Model:
                           Estimate Std. Error    z value     Pr(>|z|)
(Intercept)             -9.18954044 0.70562425 -13.023278 9.021385e-39
Pregnancies              0.14271089 0.02753313   5.183244 2.180599e-07
Glucose                  0.03690915 0.00348911  10.578386 3.753703e-26
BMI                      0.08861285 0.01471174   6.023276 1.709217e-09
DiabetesPedigreeFunction 0.89287591 0.29527948   3.023833 2.495940e-03
```

FIGURE 41 – Modèle avec les variables minimisant le critère du BIC

Autres modèles

— Nous créons un modèle similaire au modèle complet mais contenant la variable `log(Insulin)` au lieu de `Insulin`. En effet nous avions remarqué dans les boxplots que l'étendue de cette variable était bien plus importante que les autres.

— Un modèle similaire à celui des variables significatives en les remplaçant quand cela est possible par la variable qui leur est corrélée. C'est-à-dire remplacer la variable `Pregnancies` par `Age`, la variable `BMI` par `SkinThickness` puis enfin la variable `Glucose` par `Insulin`.

— Un modèle similaire au modèle complet découpant la variable `Age` en classes d'âge (en utilisant les quartiles de cette variable).

— Un modèle similaire au modèle complet découpant la variable `Pregnancies` en différents groupes (en utilisant les quartiles).

3.3.2 Recherche du modèle optimal

La recherche du modèle optimal se fait par un algorithme similaire à celui réalisé en analyse discriminante. C'est-à-dire que nous partitionnons à chaque itération de la boucle le jeu de données en un échantillon d'apprentissage et un échantillon test.

A chaque répétition, l'algorithme classera les modèles du meilleur au moins bon, par rapport à la capacité du modèle à minimiser le taux de faux négatif.

La matrice 'Scores A' rendra un score d'apparition pour chaque modèle. On retient alors celui ayant le score le plus élévé.

```
N = 1000
scores_A = rep(0,7)


for (k in 1:N){
  sample = sample.split(Outcome, SplitRatio = 0.8)
  train = subset(diabete_new, sample == TRUE)
  test = subset(diabete_new, sample == FALSE)

  modele_1 = glm(Outcome ~ . - classe_age - classe_preg , data = train, family = binomial(link = "logit"))
  modele_2 = glm(Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction , data = train, family=binomial(link="logit"))
  modele_3 = glm(Outcome ~ Age + Glucose + BMI + DiabetesPedigreeFunction , data = train, family=binomial(link="logit"))
  modele_4 = glm(Outcome ~ Age + Insulin + SkinThickness + DiabetesPedigreeFunction , data = train, family=binomial(link="logit"))
  modele_5 = glm(Outcome ~ Pregnancies + Glucose + log(Insulin) + SkinThickness + BMI + DiabetesPedigreeFunction + Age, data = train, family=binomial(link="logit"))
  modele_6 = glm(Outcome ~ Pregnancies + Glucose + Insulin + SkinThickness + BMI + DiabetesPedigreeFunction + classe_age, data = train, family=binomial(link="logit"))
  modele_7 = glm(Outcome ~ classe_preg + Glucose + Insulin + SkinThickness + BMI + DiabetesPedigreeFunction + Age, data = train, family=binomial(link="logit"))

  liste_modeles = list(modele_1, modele_2, modele_3, modele_4, modele_5, modele_6, modele_7)
  n = length(liste_modeles)
  erreur = 1
  j = 0
  A = matrix(0, nrow = 2, ncol=n)
  A[1,]= 1:n
```

```
  for (i in 1:n){
    outcome.pred = predict(liste_modeles[[i]], newdata=test, type="response")
    erreur_pred = prop.table(table(outcome.pred>0.35, test$Outcome))[2]
    A[2,i] = erreur_pred
 }
  }
  A_tri = A[,order(A[2,], decreasing = FALSE)]
  for (i in 1:n){
    scores_A[A_tri[1,i]] = scores_A[A_tri[1,i]] + (11-i)}

}
scores_A
```

Étant dans un contexte médical notre but est de minimiser le taux de faux négatifs, c'est à dire le nombre de patientes déclarées comme saines alors qu'elles sont atteintes du diabète, tout en gardant un taux de faux positifs raisonnable.

En effet, on ne souhaite pas non plus un taux de faux positifs trop élevé. Dans un contexte médical, on veut éviter de considérer trop de patientes comme étant diabétiques alors qu'elles ne le sont pas, cela n'est pas souhaitable d'un point de vue financier.

Nous choisissons tout d'abord un seuil égal à 0.5, celui ci sera éventuellement changé après l'observation des histogramme des scores.

Modèle final Le modèle retenu est

$$\text{logit}(\pi_\beta(x)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_6 X_6 + \beta_7 X_7$$

Soit contenant les variables `Pregnancies`, `Glucose`, `BMI` et `DiabetesPedigreeFunction`.

Avec notre seuil de 0.5 On obtient la matrice de confusion suivante

	$Outcome = 0$	$Outcome = 1$
$Y_{pred} = 0$	84	27
$Y_{pred} = 1$	16	27

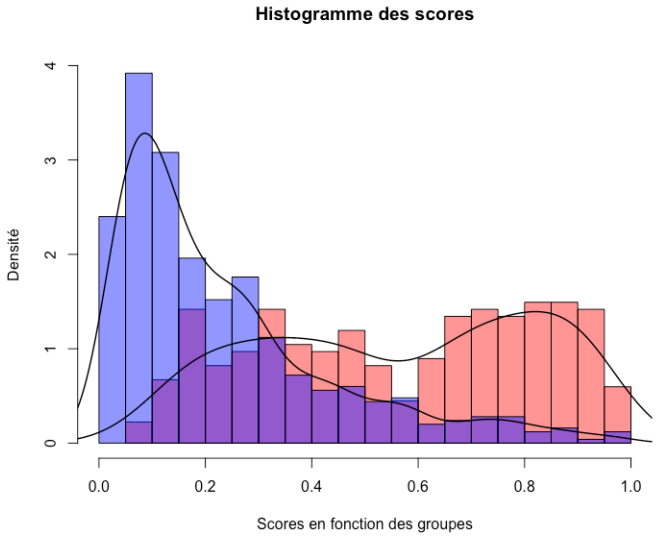


FIGURE 42 – Histogramme des scores

En rouge, sont représentées les vrais positifs et en bleu les vrais négatifs. La courbe des vrais positifs soit les individus présentant un diabète n'est pas satisfaisante, et conduit donc à des difficultés d'interprétation.

D'après les histogrammes des scores ci dessus, le seuil optimal afin de minimiser les faux négatifs se situe à 0.35. Nous allons ainsi relancer une nouvelle fois l'algorithme avec un seuil de 0.35.

Quand nous relançons l'algorithme nous obtenons exactement le même modèle final avec une nouvelle-matrice de confusion :

avec un seuil de 0.35

	$Outcome = 0$	$Outcome = 1$
$Y_{pred} = 0$	78	12
$Y_{pred} = 1$	22	42

avec un seuil de 0.30

	$Outcome = 0$	$Outcome = 1$
$Y_{pred} = 0$	70	9
$Y_{pred} = 1$	30	45

Avec un seuil = 0.35 :
Le taux d'individus sains bien classés est de 78 %
Le taux d'individus malades bien classés est de 77.78%
Soit un taux de bon classement près de 78 %

Avec un seuil = 0.30 :
Le taux d'individus sains bien classés est de 70 %
Le taux d'individus malades bien classés est de 83 %
Soit un taux de bon classement près de 75 %

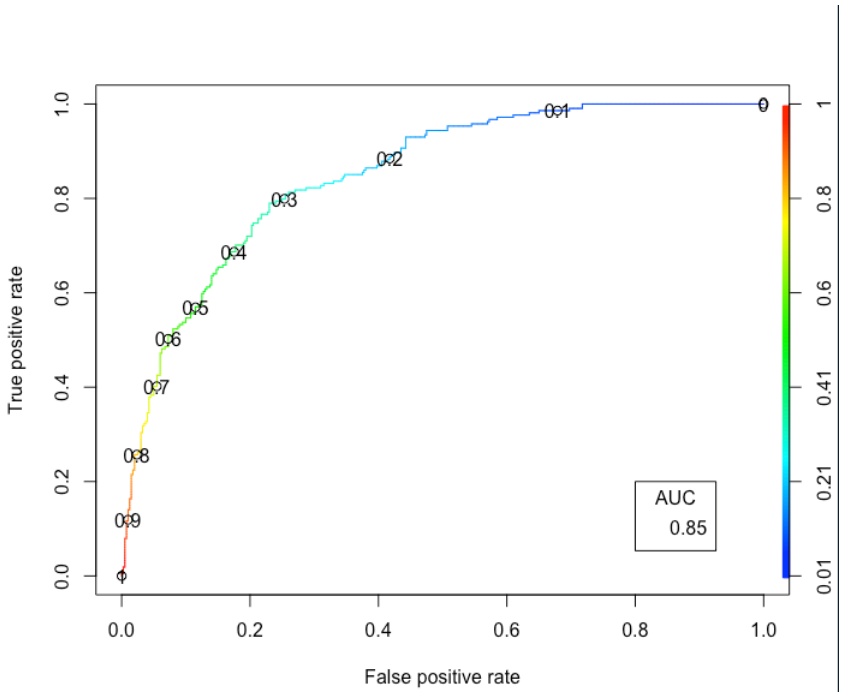


FIGURE 43 – Courbe ROC du modèle final

La courbe ROC tend vers le coin supérieur gauche et on a une aire sous la courbe égale à 0.85, ainsi le modèle retenu (celui contenant les variables significatives) a un taux de bon classement de 75%, tout cela en gardant 4 variables explicatives sur les 8 initiales.

Remarque 4. Dans un contexte médical, ces résultats ne sont pas très satisfaisants. On ne peut pas se permettre d'avoir 1 patient sur 4 qui est mal classé.
Nous pouvons émettre plusieurs hypothèses quant à ces résultats.
La base de données n'est pas assez satisfaisante. Il aurait peut être fallu pour traiter les données manquantes supprimer les variables `SkinThickness` et `Insulin`.

Quatrième partie

Annexe

1 Explication des différents ratios de la base de données farms

EBITDA = "earnings before interest, taxes, depreciation, and amortization" = excédent brut d'exploitation

Capitalisation

$$r1 = \frac{\text{Dette totale}}{\text{Totalité des actifs}} \qquad r2 = \frac{\text{Capitaux propres}}{\text{Capitaux permanents}} \qquad r3 = \frac{\text{Dette de court terme}}{\text{Dette totale}}$$

$$r4 = \frac{\text{Dette de court terme}}{\text{Totalité des actifs}} \qquad r5 = \frac{\text{Dette de long et moyen terme}}{\text{Totalité des actifs}}$$

Poids des dettes

$$r6 = \frac{\text{Dette totale}}{\text{Produit brut}} \qquad r7 = \frac{\text{Dette à long et moyen termes}}{\text{Produit Brut}} \qquad r8 = \frac{\text{Dette de court terme}}{\text{Produit brut}}$$

Liquidité

$$r11 = \frac{\text{Fonds de roulement}}{\text{Produit brut}} \qquad r12 = \frac{\text{Fond de roulement}}{\text{Entrées réelles - frais financiers}} \qquad r14 = \frac{\text{Dette de court terme}}{\text{Actifs en circulation}}$$

Debt servicing

$$r17 = \frac{\text{Frais financiers}}{\text{Dette totale}} \qquad r18 = \frac{\text{Frais financiers}}{\text{Produit brut}}$$

$$r19 = \frac{\text{Frais financiers + remboursement du capital à long et moyen terme}}{\text{Produit brut}}$$

$$r21 = \frac{\text{Frais financiers}}{\text{EBITDA}}$$

$$r22 = \frac{\text{Frais financiers + remboursement du capital à long et moyen terme}}{\text{EBITDA}}$$

Capital profitability

$$r24 = \frac{\text{EBITDA}}{\text{Totalité des actifs}}$$

Earnings

$$r28 = \frac{\text{EBITDA}}{\text{Produit brut}} \qquad r30 = \frac{\text{Revenu disponible}}{\text{Produit brut}} \qquad r32 = \frac{\text{EBITDA - Frais financiers}}{\text{Produit brut}}$$

Productive activity

$$r36 = \frac{\text{Actifs immobilisés}}{\text{Produit brut}} \qquad r37 = \frac{\text{Porduit brut}}{\text{Totalité des actifs}}$$

Cinquième partie

Conclusion Générale

Nous avons dans ce rapport pu mettre en place plusieurs techniques de scoring sur des jeux de données différents. La partie affectation de score à un individu a été réalisée soit par analyse en composantes principales (ACP), soit par analyse factorielle discriminante (AFD), soit par régression logistique.

Sur l'exemple économique des exploitations agricoles, l'ACP donne une règle de classement facile à mettre en place et donnant un taux de bon classement très acceptable.

L'AFD et la régression logistique donnent elles des performances quasi-égales.

Or, comme il a déjà été mentionné plus tôt dans ce rapport, le scoring ne s'arrête pas à une simple classification. Une fois le modèle optimal déterminé par une des méthodes présentées ici (AFD, régression logistique) ou d'autres méthodes de scoring, il faut choisir une valeur seuil (pivot) permettant de répondre à certaines exigences financières ou humaines. Cela est possible en observant les histogrammes de score ou les courbes ROC.

Pour l'exemple médical, il est tout à fait concevable de vouloir un taux de faux négatif inférieur à 5% pour ne pas laisser passer trop de personnes atteintes mais la contrainte financière peut imposer un taux de faux positif de 10% maximum (pour na pas tester à tord trop de personnes non malade et ainsi perdre du temps et de l'argent), auquel cas il faudra choisir un seuil répondant à ces 2 critères. (Donner les chiffres si possible).

Nous voyons donc que le scoring est un outil puissant d'aide à la décision et de détection (maladies, fraudes, etc...). Le scoring est notamment largement utilisé dans le domaine médical et financier (banques, assurances, ...) et se répand de plus en plus dans le milieu marketing afin d'augmenter les ventes et d'optimiser le ciblage de clientèle.

bibliographie$_s$*coring*