SDMA - TP3
Analyse discriminante
Régression logistique

Maha ELBAYAD

8 Décembre 2015

Analyse Discriminante

Données de l'étude:

Base de données **Iris**:

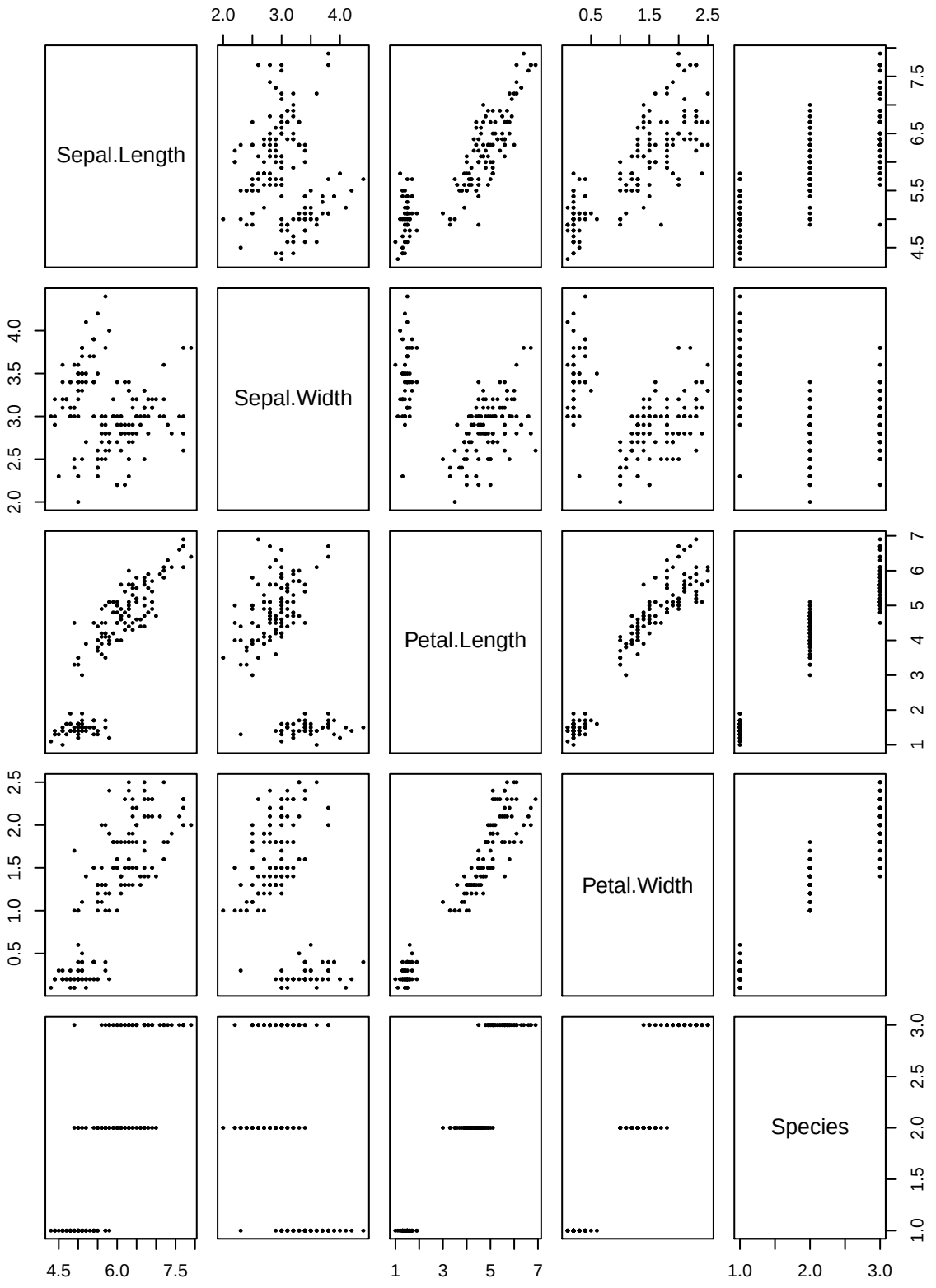
- Taille de l'échantillon: 150
- Nombre de covariables: 5

Covariables numériques:

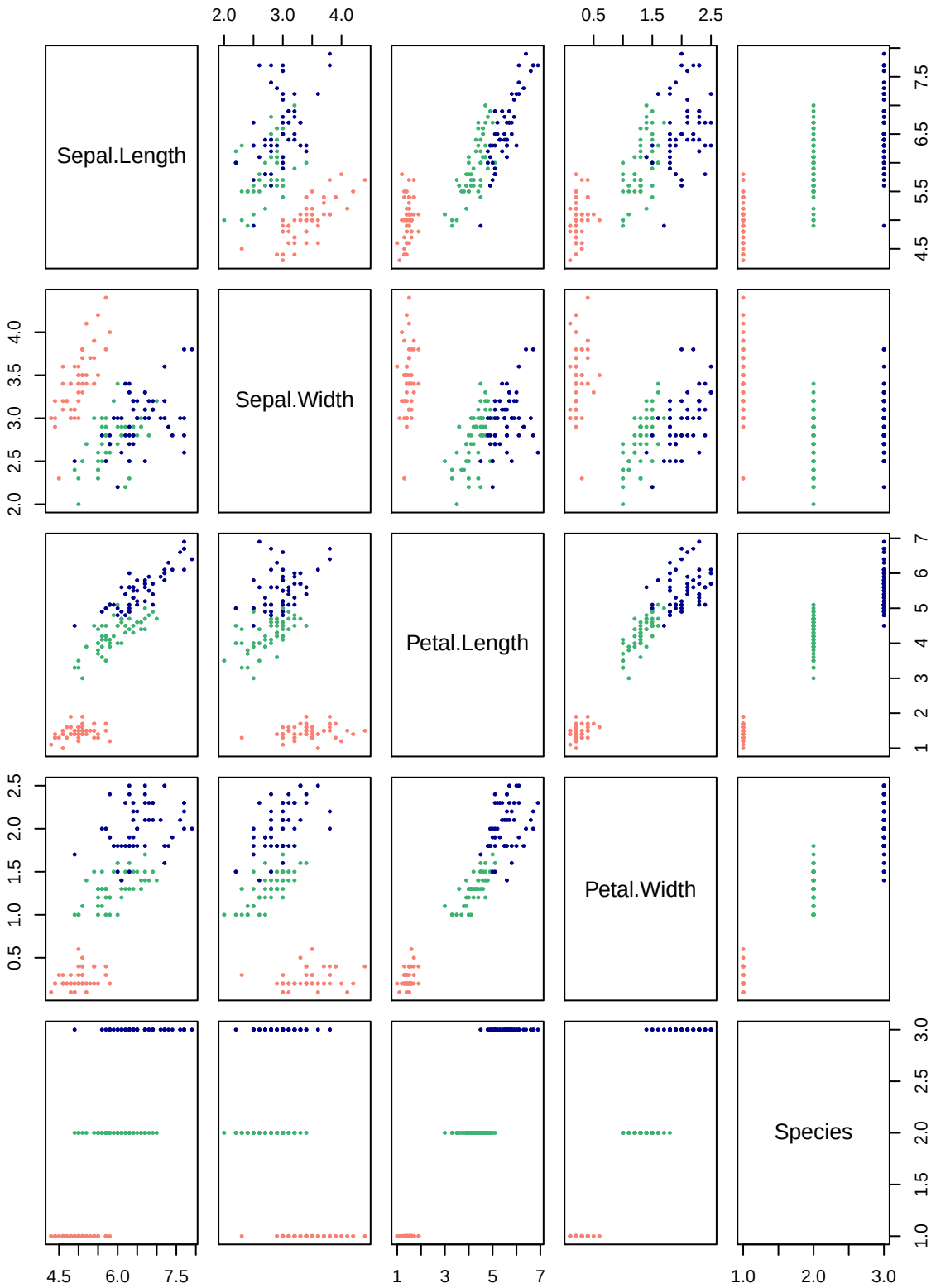
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300
Median :5.800	Median :3.000	Median :4.350	Median :1.300
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500

Moyenne des covariables par espèce:

	Effectif	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	50.00	5.01	3.43	1.46	0.25
versicolor	50.00	5.94	2.77	4.26	1.33
virginica	50.00	6.59	2.97	5.55	2.03


Iris
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species

Iris par espèce



Analyse discriminante linéaire prédictive (ADL) (librairie MASS)

Prior:

setosa	versicolor	virginica
0.33	0.33	0.33

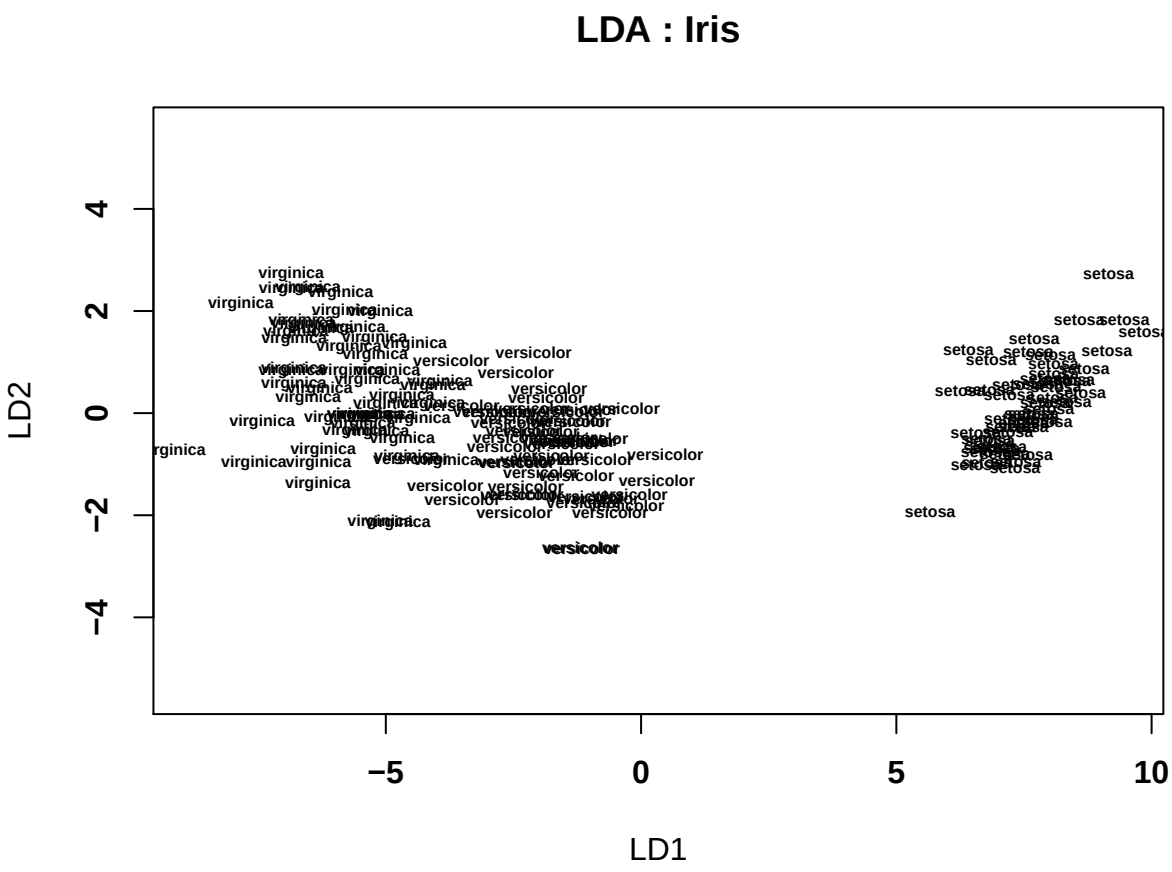
Comme on a autant d'exemples dans chaque espèce le prior est équiréparti.

Règles de décision:

	LD1	LD2
Sepal.Length	0.83	0.02
Sepal.Width	1.53	2.16
Petal.Length	-2.20	-0.93
Petal.Width	-2.81	2.84

La frontière de décision *LD1* permet de séparer la classe *setosa* des deux autres classes. Elle a une forte trace de 0.9912. Alors que *LD2* est peu performante.

La fonction plot de la classe *LDA* affiche la projection des samples sur les deux frontières de décision (LD1,LD2)



##Apprentissage/Test:

Matrice de confusion -1:

Avec 1:setosa 2:versicolor 3:virginica

	pred 1	pred 2	pred 3	Erreur
true 1	8	0	0	0.00
true 2	0	9	0	0.00
true 3	0	0	13	0.00

L'erreur sur la base de test est de 0% contre 2.5% sur la base d'apprentissage.

Matrice de confusion -2:

	pred 1	pred 2	pred 3	Erreur
true 1	6	0	0	0.00
true 2	0	12	1	0.08
true 3	0	1	10	0.09

L'erreur sur la base de test est de 6.67% contre 1.67% sur la base d'apprentissage.

Pour cette répartition train/test l'erreur vient surtout de la confusion entre *versicolor* et *virginica*.

On remarque que l'erreur du modèle dépend de la répartition aléatoire train/test. Comme la base Iris est de petite taille, il faut réitérer cette procédure ou estimer l'erreur par ré-échantillonnage pour une meilleure évaluation de la performance du LDA.

Analyse discriminante quadratique prédictive (ADQ) (librairie MASS)

QDA - 1ère répartition:

L'erreur sur la base de test est de 0% contre 2.5% sur la base d'apprentissage.

QDA - 2ème répartition:

L'erreur sur la base de test est de 3.33% contre 1.67% sur la base d'apprentissage.

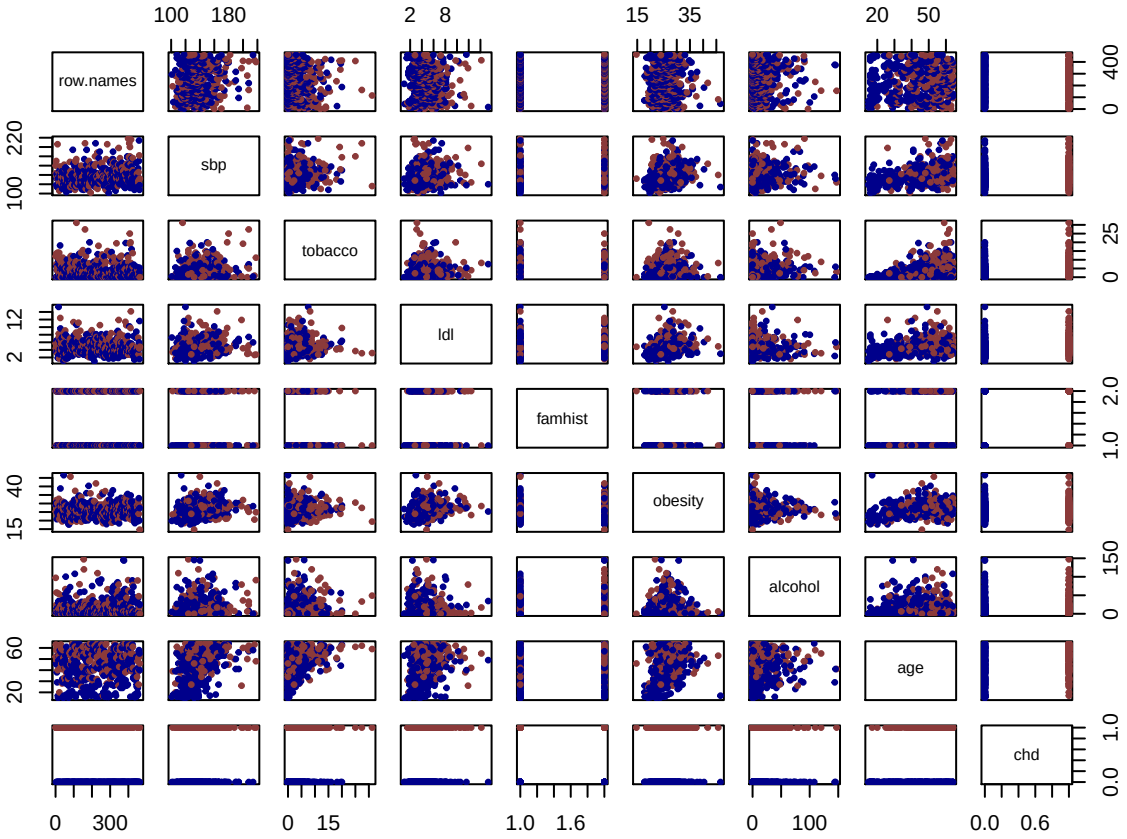
Pour mieux comparer LDA et QDA, on réitère 50 fois la procédure de répartition train/test:

- E(erreur-train(LDA))=2.02%
- E(erreur-train(QDA))=1.83%
- E(erreur-test(LDA))=2%
- E(erreur-test(QDA))=2.67%

On remarque que le QDA sur-apprend la base d'apprentissage et a donc une erreur sur la base de test légèrement supérieure à celle du LDA.

Régression logistique:

- sbp: pression systolique
- tobacco: consommation de tabac cumulative
- ldl: cholestérol
- famhist: antécédents familiaux de problèmes cardiaques (catégorique)
- obesity : obésité
- alcohol: Consommation d'alcool
- age: Age
- chd: Présence de maladie coronarienne (cible)



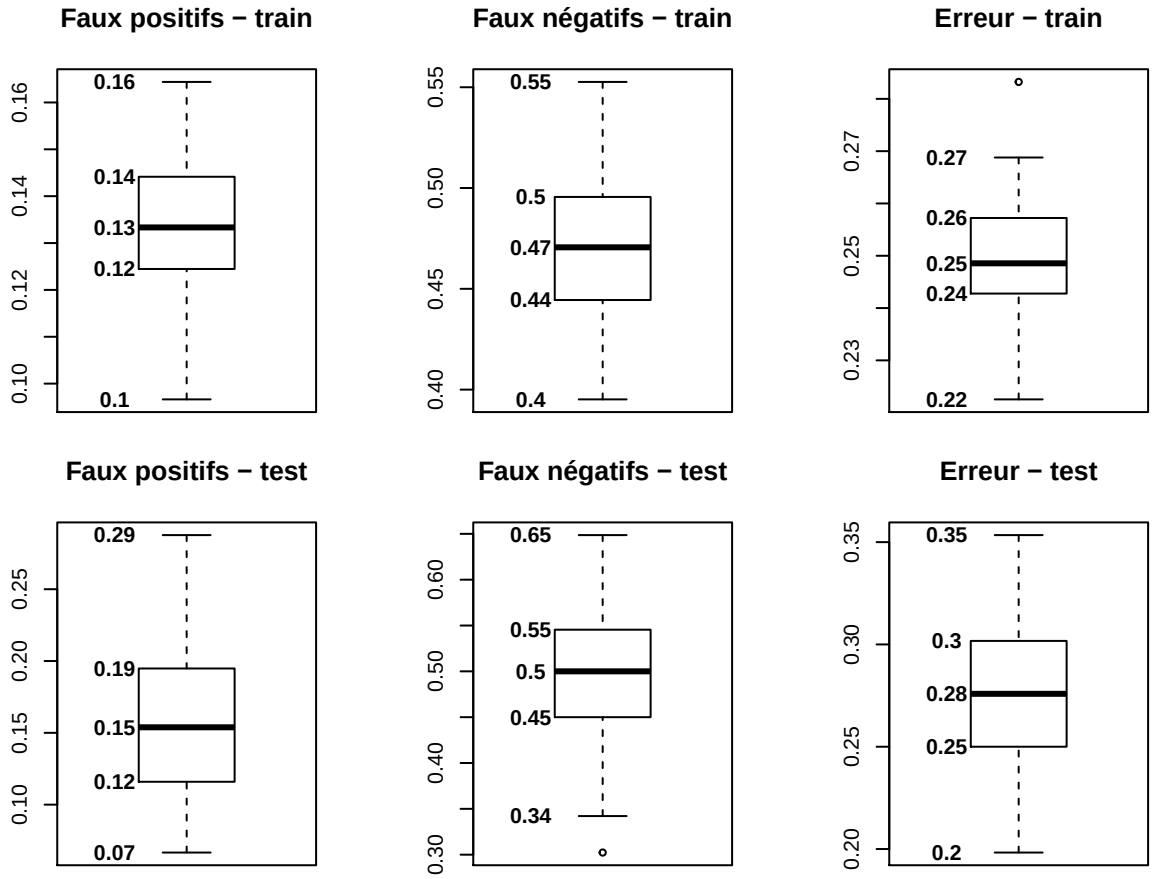
Matrise de confusion - base d'apprentissage:

	pred 0	pred 1
true 0	267	35
true 1	77	83

Le taux de faux positifs vaut 11.6% et le taux de faux négatifs 48.1%

Ce modèle n'arrive pas à détecter plus de 48% de malades (faux négatifs), ce qui a le plus d'importance dans un contexte médical. Les faux positifs quant à eux, ne présentent pas autant de risque.

Validation croisée:



L'erreur moyenne sur la base de test (25%) est de 27.7%. Avec la réitération de la procédure de sélection de la base d'apprentissage/test on estime rigoureusement l'espérance de l'erreur et on peut également évaluer la stabilité de modèle via la variance de l'erreur.

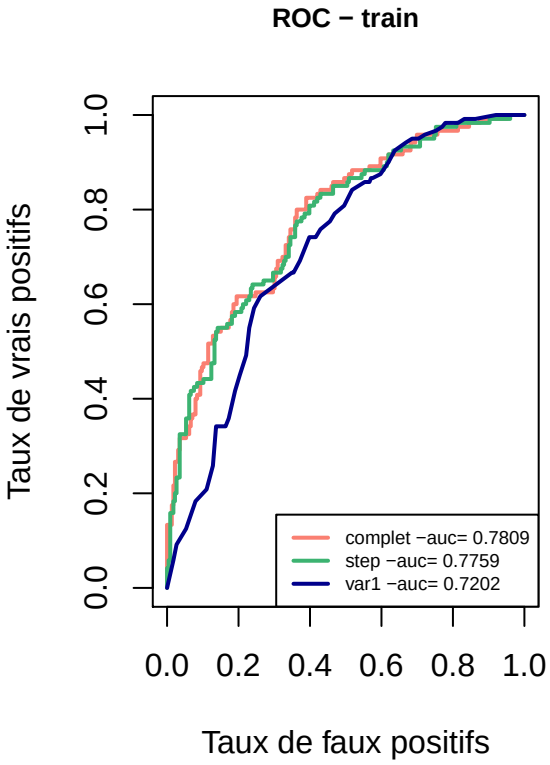
Régression logistique avec sélection de variables:

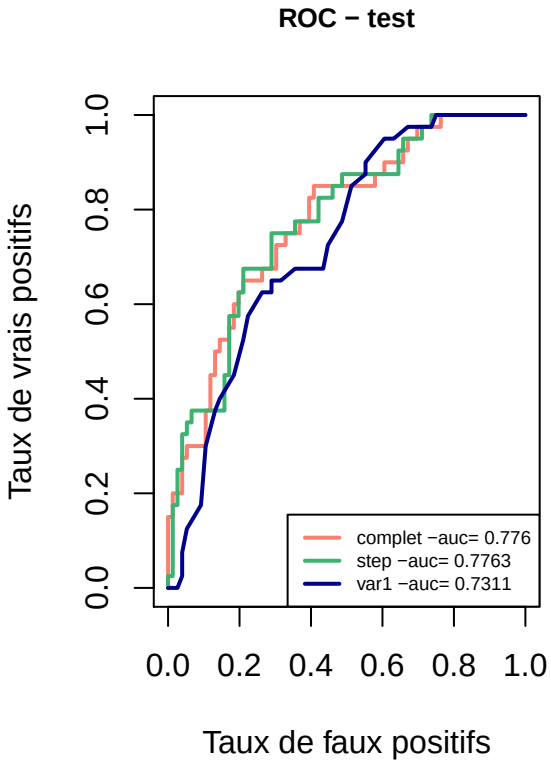
Avec *step* dans les deux directions, les covariables les plus significatives sélectionnées dans l'ordre croissant d'importance sont: age -> famhist -> tobacco -> ldl. On conclut que l'âge et les antécédents familaux prédominent les covariables diététiques (tabac, alcool..)

	Estimate	Pr(>|z|)	signif
(Intercept)	-4.2	3.26e-17	***
age	0.044	6.17e-06	***
famhistPresent	0.924	3.46e-05	***
tobacco	0.0807	0.00156	**
ldl	0.168	0.00198	**

Table 1: Coefficients estimés du modèle sélectionné

Les courbes ROC:





On note que le sous-modèle (chd ~ age+famhist+tobacco+ld) a une performance très proche de celle du modèle complet sur la base d'apprentissage et arrive même à le surperformer sur la base de test ($+3.29 \times 10^{-4}$ de plus en ROC AUC). Le modèle *var1* à une variable (chd ~ âge) n'est pas loin derrière et pourrait être suffisant selon l'usage.