1 Objectif

Equivalences entre l'analyse discriminante linéaire et la régression linéaire multiple.

Bien que s'inscrivant toutes deux dans le cadre de l'analyse prédictive, l'analyse discriminante linéaire et la régression linéaire multiple répondent à des problématiques différentes. La première cherche à prédire une variable cible qualitative nominale à partir d'un ensemble de variables prédictives quantitatives (ou qualitatives recodées en indicatrices numériques). Pour la seconde, la variable cible est quantitative. La finalité, les calculs sous-jacents et le mécanisme inférentiel ne sont pas les mêmes.

Pourtant, de nombreux auteurs indiquent qu'il y a des similarités entre ces deux approches. Mieux même, dans le cas particulier d'une variable cible binaire, il est possible de reproduire à l'identique les sorties de l'analyse discriminante à partir des résultats de la régression (Bishop, 2007, pages 189 - 190 ; Duda et al., 2001, pages 242 – 243 ; Huberty et Olejnik, 2006, pages 353 – 355 ; Nakache et Confais, 2003, pages 14 – 16 ; Saporta, 2006, pages 451 – 452 ; Tomassone et al., 1988, pages 36 – 38). Malheureusement, si les références consultées montrent effectivement les connexions entre les expressions matricielles, certaines explicitant les formules de transition, aucune ne détaille les calculs sur un exemple numérique, rendant la démonstration par trop abstraite. Le lecteur a du mal à percevoir la portée réelle de cette équivalence. A force de chercher sur le web (en français et en anglais), j'ai fini par trouver un exemple traité qui met en évidence la relation. Les coefficients des fonctions linéaires issues des deux approches sont proportionnels, hélas sans que l'auteur n'indique l'expression mathématique du rapport entre les coefficients (Desbois, 2003 ; page 31).

Ce tutoriel reprend la même idée. A partir du traitement d'un fichier de données, nous décrivons les connexions entre les deux approches dans le cas d'une variable cible binaire. Nous détaillons les formules permettant de retrouver les coefficients de l'analyse discriminante à partir de ceux de la régression linéaire. Il apparaît que, si l'équivalence est totale lorsque les classes sont équilibrées, il est nécessaire d'introduire une correction additionnelle de la constante lorsque les effectifs ne sont pas identiques dans les deux groupes (Hastie et al, 2009, page 110). La formule correspondante – introuvable également dans les différentes références – est explicitée. Nous réalisons les calculs sous **Tanagra** (classes équilibrées) et sous **R** (classes non équilibrées). Nous nous appuyons essentiellement sur la référence Tomassone et al. (1988), ouvrage absolument remarquable sur ce sujet, mais aussi de manière générale sur les différents thèmes qui y sont développés.

2 Traitement des classes équilibrées

2.1 Données – IRIS

Nous utilisons un sous-ensemble du fichier IRIS[1] dans cette partie. Nous ne conservons que les **p = 2** dernières variables prédictives {petal-length, petal-width}, et les observations correspondants aux **K = 2** groupes {iris-versicolor, iris-virginica}, soit **n = 100** observations. Nous avons adjoint au fichier la variable « **y** » dont nous préciserons la nature plus loin.

Voici les 6 premières lignes du fichier :

pet.length	pet.width	species	y
4.7	1.4	versicolor	0.5
4.5	1.5	versicolor	0.5
4.9	1.5	versicolor	0.5
4.0	1.3	versicolor	0.5
4.6	1.5	versicolor	0.5
4.5	1.3	versicolor	0.5

Figure 1 - Premières lignes du tableau de données - IRIS binaire

Puisque nous n'avons que 2 variables, il est aisé de représenter les observations dans le plan en les différenciant selon leur classe d'appartenance.

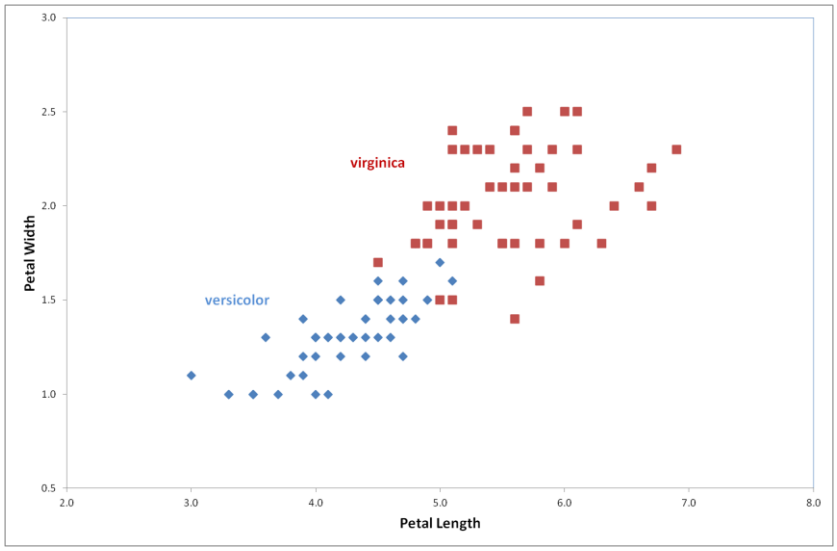


Figure 2 - Nuages de points conditionnellement aux classes

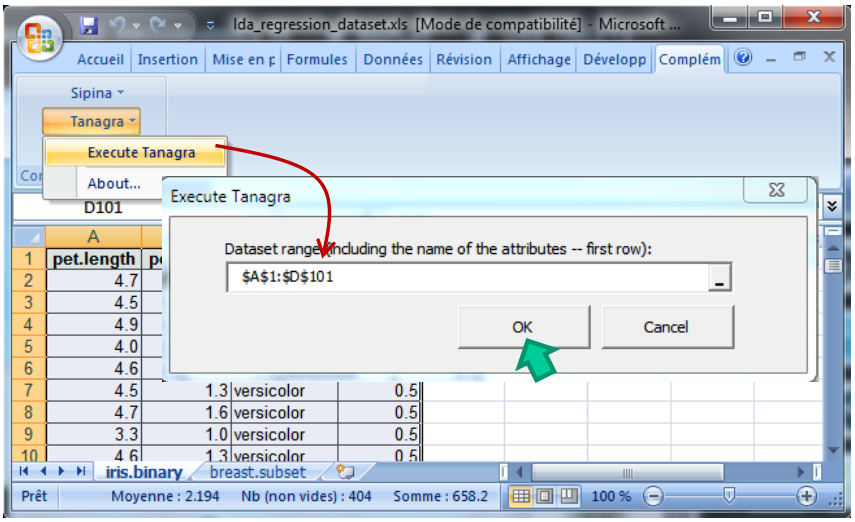
Les 2 groupes d'individus sont relativement distincts. Trouver une frontière linéaire qui permet de les séparer sera aisé. Le taux d'erreur du modèle sera vraisemblablement faible. Les individus mal classés seront localisés dans les parties recouvrantes des nuages conditionnels.

[1] http://archive.ics.uci.edu/ml/datasets/Iris ; un fichier fameux, incontournable dans notre domaine. Cf. R.A. Fisher, « The use of Multiple Measurements in Taxonomic Problems », in Annals of Eugenics, 7(2), 179-188, Septembre 1936.

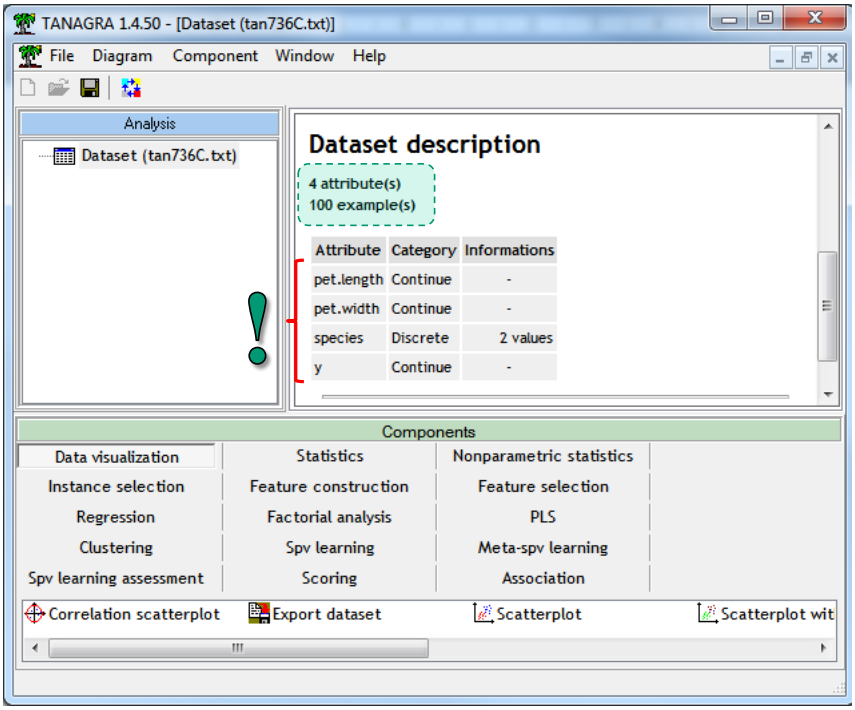
2.2 Analyse discriminante avec Tanagra – Lecture des résultats

2.2.1 Importation des données

Nous souhaitons réaliser une analyse discriminante prédictive avec Tanagra. Nous sélectionnons l'ensemble de données sous Excel et nous l'envoyons à Tanagra via la macro-complémentaire « tanagra.xla »[2].



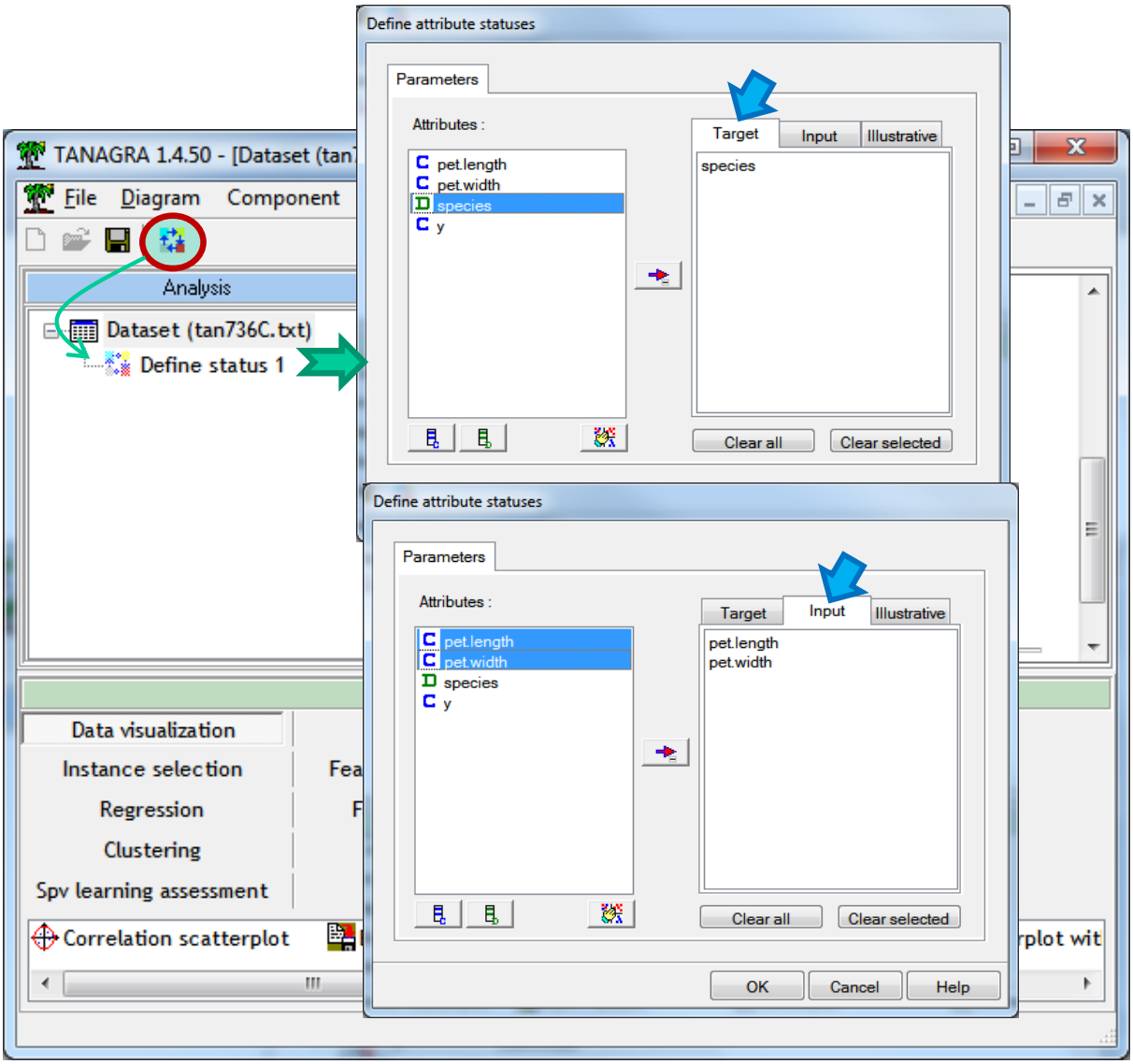
Tanagra est automatiquement démarré : 4 colonnes sont importées, avec 100 observations.



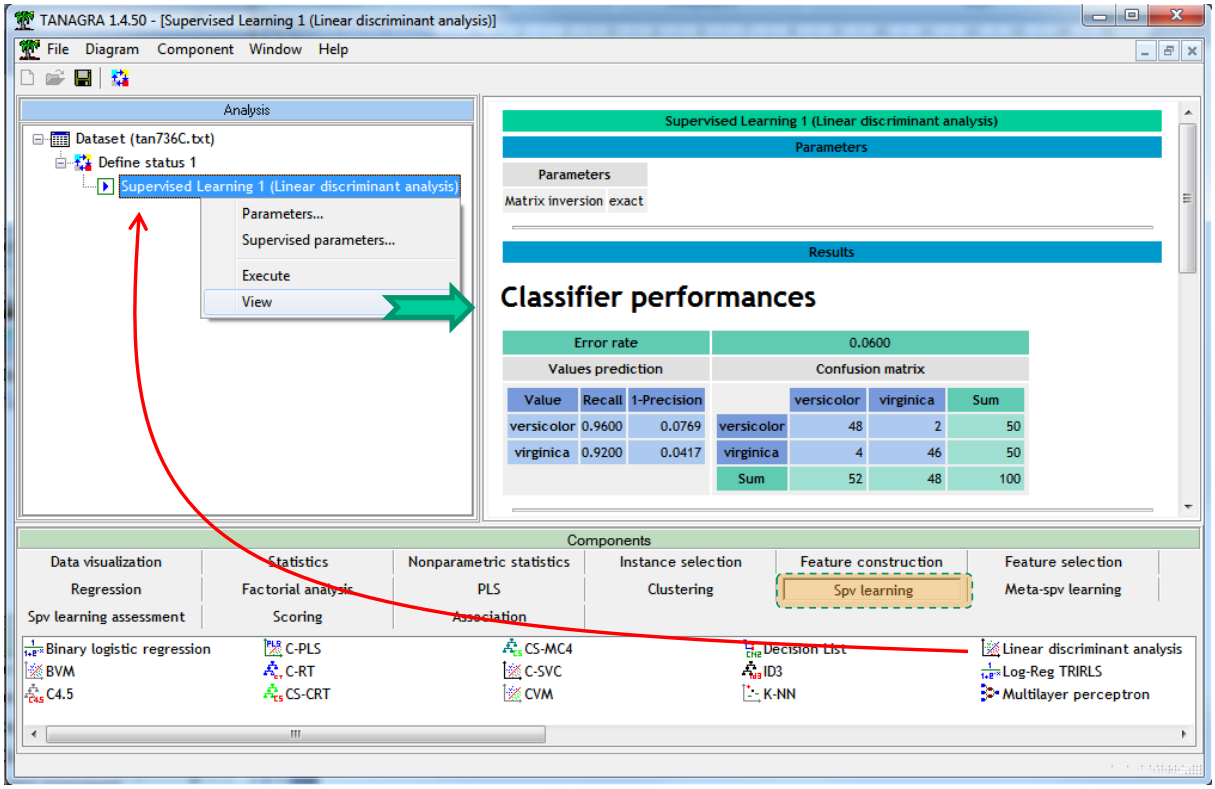
[2] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour l'installation et l'utilisation de la macro-complémentaire. Ce type de dispositif existe également pour le tableur Calc des suites « Libre Office » et « Open Office » (http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html).

2.2.2 Analyse discriminante

Pour réaliser l'analyse discriminante, nous devons tout d'abord définir le rôle des variables. Nous utilisons le composant DEFINE STATUS en cliquant sur le raccourci dans la barre d'outils. Nous plaçons SPECIES en TARGET, PET.LENGTH et PET.WIDTH en INPUT. La variable Y n'est pas utilisée à ce stade.



Nous ajoutons l'outil LINEAR DISCRIMINANT ANALYSIS (onglet SPV LEARNING) dans le diagramme. Nous actionnons le menu contextuel VIEW pour accéder aux résultats.



2.2.3 Lecture des résultats

Matrice de confusion. La partie « Classifier performances » inclut la matrice de confusion calculée sur l'échantillon d'apprentissage.

Classifier performances

Error rate			0.06			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		versicolor	virginica	Sum
versicolor	0.96	0.0769	versicolor	48	2	50
virginica	0.92	0.0417	virginica	4	46	50
			Sum	52	48	100

Figure 3 - Matrice de confusion

Les groupes (G) sont parfaitement équilibrés avec $\mathbf{n_1 = 50}$ « G_1 : versicolor » et $\mathbf{n_2 = 50}$ « G_2 : virginica ». 6 observations sont mal classées (taux d'erreur 6 / 100 = 6%) avec 4 observations « virginica » étiquetées « versicolor », et 2 inversement. Nous les visualiserons lorsque nous tracerons la frontière séparant les classes dans l'espace de représentation (Figure 7).

MANOVA. L'analyse de variance multidimensionnelle correspond à un test de comparaison des barycentres conditionnels. Le Λ de Wilks est le rapport entre les variances généralisées intra-classes et totales. Plus il se rapproche de 0, plus les barycentres sont distincts. Dans notre cas, nous avons $\Lambda = 0.2802$, ce qui laisse augurer d'une bonne séparabilité des groupes,

confirmée d'une part, par le graphique des nuages de points conditionnels (Figure 2), d'autre part, par le faible taux d'erreur (Figure 3).

MANOVA

Stat	Value	p-value
Wilks' Lambda	0.2802	-
Bartlett -- C(2)	123.3935	0
Rao -- F(2, 97)	124.5641	0

Figure 4 - Evaluation globale du modèle - Analyse discriminante

Le Λ de Wilks peut s'appliquer pour un nombre quelconque de groupes ($K \geq 2$). Pour le cas binaire ($K = 2$), on peut en déduire la distance entre les centres de classes μ_1 (versicolor) et μ_2 (virginica), appelée « distance de Mahalanobis » (D^2). Par rapport à la distance euclidienne habituelle, elle a pour particularité de tenir compte de la forme des nuages de points,

$$D^2 = \frac{1-\Lambda}{\Lambda} \times \frac{n(n-2)}{n_1 \times n_2}$$

Dans notre cas,

$$D^2 = \frac{1-0.2802}{0.2802} \times \frac{100(98)}{50 \times 50} = 10.0678$$

Graphiquement, nous situons les centres de classes (μ_1, μ_2) – de coordonnées μ_1 = (4.26, 1.33) et μ_2 = (5.55, 2.03) – et leur écartement D^2 (Figure 5).

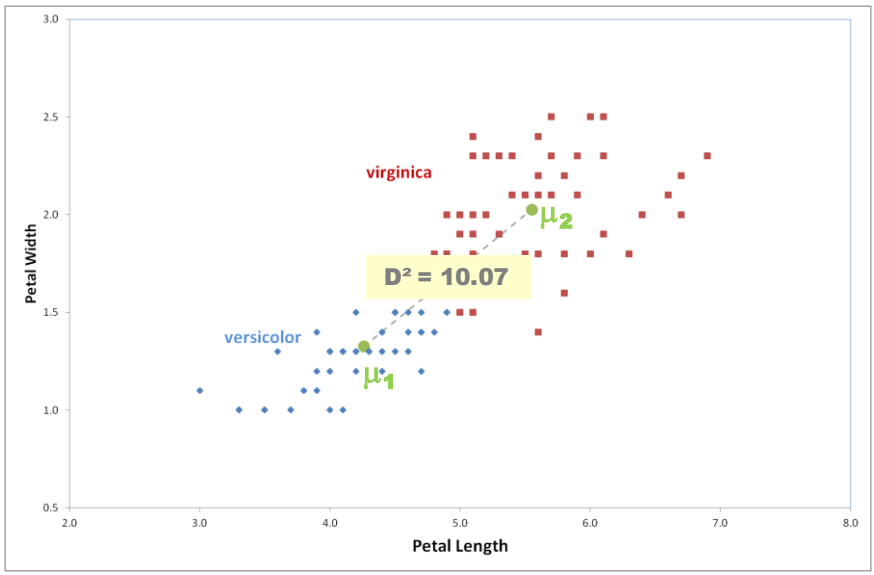


Figure 5 - Barycentres conditionnels - Distance de Mahalanobis (D^2)

Pour tester la significativité de l'écartement, nous utilisons la statistique F de Rao qui suit une loi de Fisher sous l'hypothèse nulle d'égalité des moyennes conditionnelles[3]. Pour le fichier

[3] R. Rakotomalala, « Comparaison de populations – Tests paramétriques », version 1.2, Mai 2010 ; pages 87 et 88.

IRIS, **F = 124.5641**, la statistique suit une loi de Fisher à (2, 97) degrés de liberté sous H_0 : nous rejetons l'hypothèse nulle au risque 5%.

Fonctions de classement – **Fonction score**. Les fonctions de classement permettent d'affecter les groupes aux individus à partir de leur description. Nous avons autant de fonctions que de groupes (Huberty et Olejnik, 2006 ; page 274).

$$D(G_1, X) = a_0 + a_1 * X_1 + a2 * X_2$$

$$D(G_2, X) = b_0 + b_1 * X_1 + b2 * X_2$$

Dans le cas binaire (K = 2), nous pouvons déduire la fonction score qui est formée à partir de la différence termes à termes des coefficients des fonctions de classement. Elle renvoie une valeur proportionnelle à la probabilité d'appartenir à G_1. Parmi les classifieurs linéaires, c'est une alternative possible à l'équation LOGIT fournie par la régression logistique[4].

$$D(X) = \theta_0 + \theta_1 * X_1 + \theta_2 * X_2$$

Avec

$$\theta_j = (a_j - b_j)$$

Tanagra fournit les fonctions de classement, nous en avons déduit la fonction score.

	Classification functions		Score function
Attribute	**versicolor**	**virginica**	**D(X)**
pet.length	**14.40029**	**17.164859**	**-2.764569**
pet.width	**7.824622**	**17.104674**	**-9.280052**
constant	**-36.55349**	**-65.66983**	**29.116340**

Figure 6 - Fonctions de classement et fonction score

La règle d'affectation pour un individu ω à classer s'écrit :

Si $D[X(\omega)] \geq 0$ **Alors** Versicolor **Sinon** Virginica

Ainsi, pour un individu (pet.length = 4.7, pet.width = 1.4), nous calculons :

$$D = 29.116340 + (-2.764569 * 4.7) + (-9.280052 * 1.4) = 3.13 > 0$$

Nous lui attribuons la classe « versicolor », ce qui paraît sensé au regard de sa localisation géographique (Figure 7).

Frontière entre les classes. D(X) = 0 définit la frontière induite par la méthode pour séparer les classes dans l'espace de représentation. Dans le plan, nous pouvons la représenter par une droite (Figure 7).

[4] Tutoriels Tanagra, « Classifieurs linéaires », http://tutoriels-data-mining.blogspot.fr/2013/05/classifieurs-lineaires.html

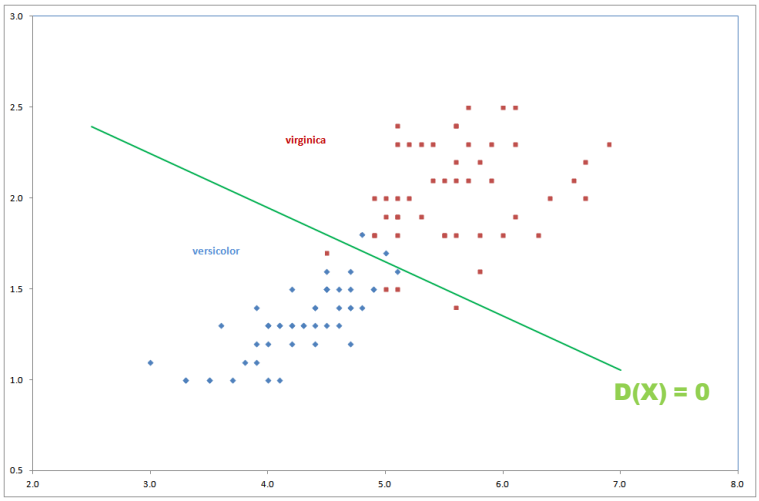


Figure 7 - Frontière induite par l'analyse discriminante linéaire

Nous distinguons parfaitement les 6 individus mal classés signalés dans la matrice de confusion (Figure 3). Il s'agit de 4 virginica en rouge (resp. 2 versicolor en bleu) placés du mauvais côté de la frontière.

Pertinence des variables prédictives. La partie « **Statistical Evaluation** » du tableau des coefficients vise à apprécier l'influence des variables prédictives dans le modèle. Plusieurs points de vues sont possibles : vérifier que les coefficients d'une variable prennent les mêmes valeurs dans l'ensemble des fonctions de classement ; tester l'égalité des moyennes conditionnelles de la variable, en tenant compte du rôle des autres descripteurs (ce n'est pas un simple test univarié donc). Concrètement, la statistique de test $\mathbf{F_j}$ est basée sur la comparaison des Λ de Wilks avec et sans la variable X_j à évaluer. Sous H0, elle suit une loi de Fisher à $(1, n - p - K + 1)$ degrés de liberté[5] [$(1, n - p - 1)$ puisque $K = 2$ dans la configuration qui nous concerne].

LDA Summary

	Classification functions		Statistical Evaluation			
Attribute	**versicolor**	**virginica**	**Wilks L.**	**Partial L.**	**F(1,97)**	**p-value**
pet.length	14.40029	17.164859	0.314202	0.89192	11.75412	0.000893
pet.width	7.824622	17.104674	0.381538	0.734509	35.06098	0.000000
constant	-36.55349	-65.66983	-			

Figure 8 - Pertinence des variables - Analyse discriminante

Nous constatons que les 2 variables sont pertinentes à 5%. Nous retiendrons en particulier les valeurs de F pour les comparer avec les statistiques de tests de significativité de la régression.

[5] R. Rakotomalala, « Analyse discriminante linéaire », Support de cours, page 10.

2.3 Comparaison avec SAS

Les mêmes résultats sont disponibles dans deux procédures différentes de SAS. La PROC DISCRIM fournit l'évaluation globale et les fonctions de classement.

```
proc discrim data = mesdata.iris_binary manova;
class species;
var pet_length pet_width;
priors proportional;
run;
```

Nous obtenons (voir Figure 4 et Figure 6).

The SAS System

The DISCRIM Procedure

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=47.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.28024304	124.56	2	97	<.0001
Pillai's Trace	0.71975696	124.56	2	97	<.0001
Hotelling-Lawley Trace	2.56833127	124.56	2	97	<.0001
Roy's Greatest Root	2.56833127	124.56	2	97	<.0001

Linear Discriminant Function for species			
Variable	Label	versicolor	virginica
Constant		-36.55349	-65.66983
pet_length	pet#length	14.40029	17.16486
pet_width	pet#width	7.82462	17.10467

La PROC STEPDISC produit les statistiques de test F_j permettant de statuer sur les contributions des variables.

```
proc stepdisc data = mesdata.iris_binary method = backward;
class species;
var pet_length pet_width;
run;
```

Nous observons les mêmes statistiques F que sous Tanagra (Figure 8).

The SAS System

The STEPDISC Procedure
Backward Elimination: Step 1

Statistics for Removal, DF = 1, 97				
Variable	Label	Partial R-Square	F Value	Pr > F
pet_length	pet#length	0.1081	11.75	0.0009
pet_width	pet#width	0.2655	35.06	<.0001

2.4 Régression linéaire pour le classement

2.4.1 Principe – Travailler avec une variable cible recodée

La régression linéaire multiple vise à prédire les valeurs d'une variable cible quantitative à partir d'un ensemble de descripteurs. Nous disposons de toute une panoplie inférentielle pour évaluer le modèle dans sa globalité et pour apprécier la contribution des variables[6].

Voyons comment réaliser une régression sur le fichier IRIS. Cela passe obligatoirement par le recodage approprié de la variable cible « species ». La nouvelle variable cible Y prend deux valeurs numériques possibles (y_1, y_2), avec pour un individu ω :

$$y(\omega) = \begin{cases} y_1 \text{ quand } \omega \in G_1 \\ y_2 \text{ quand } \omega \in G_2 \end{cases}$$

Nous obtenons une équation de régression :

$$R(X) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

Où β_j sont les coefficients du modèle.

Pour un individu à classer ω, la règle d'affectation s'écrit :

Si $R[X(\omega)] \geq \bar{y}$ **Alors** Versicolor **Sinon** Virginica

La valeur seuil $\bar{y}$ est la moyenne pondérée des codes numériques :

$$\bar{y} = \frac{n_1 \times y_1 + n_2 \times y_2}{n}$$

Tout code peut convenir pourvu que $y_1 \neq y_2$. Plusieurs options sont possibles.

- Le plus simple : ($y_1 = 1$; $y_2 = 0$). Dans ce cas, le seuil est $\bar{y} = \frac{n_1}{n}$. De fait, le seuil 0.5 correspond à un cas particulier, spécifique à la configuration où les classes sont équilibrées ($n_1 = n_2$).
- Les codes $\left(y_1 = \frac{n_2}{n}; y_2 = -\frac{n_1}{n} \right)$ (Tomassone, 1988 ; page 38) présentent l'avantage d'induire un seuil nul puisque $\bar{y} = 0$. L'équation de régression s'apparente à une fonction score dans ce cas. Nous verrons qu'elle est totalement équivalente en réalité lorsque $n_1 = n_2$.

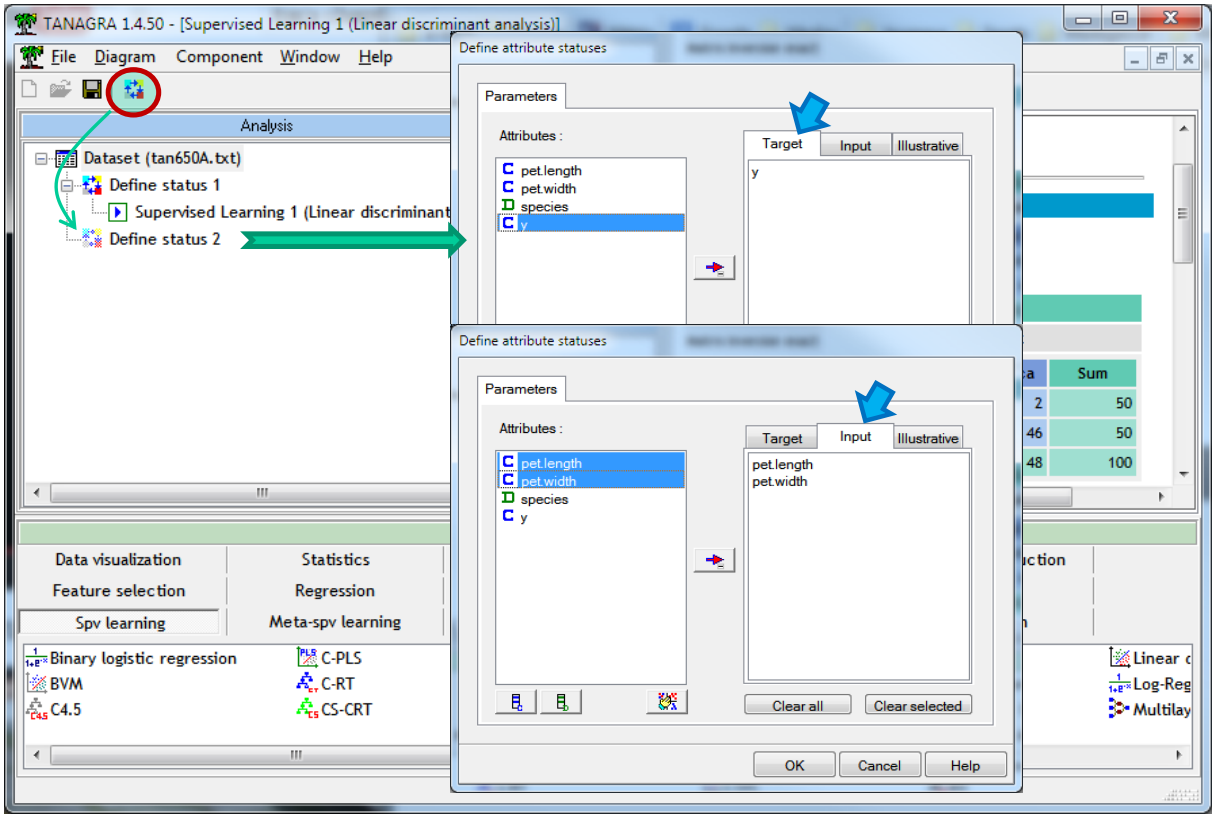
[6] Il y a pléthore de documentation sur le web. J'y ai moi-même contribué. Voir http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

- D'autres codages conduisant à une valeur seuil nulle ($\bar{y} = 0$) sont décrits dans la littérature : $\left(y_1 = \frac{n}{n_1}; y_2 = -\frac{n}{n_2} \right)$ (Duda et al., 2001, page 242 ; Saporta, 2006, page 451) ; $\left(y_1 = \sqrt{\frac{n_2}{n_1}}; y_2 = -\sqrt{\frac{n_1}{n_2}} \right)$ (Nakache et Confais, 2003 ; page 14) ; etc.

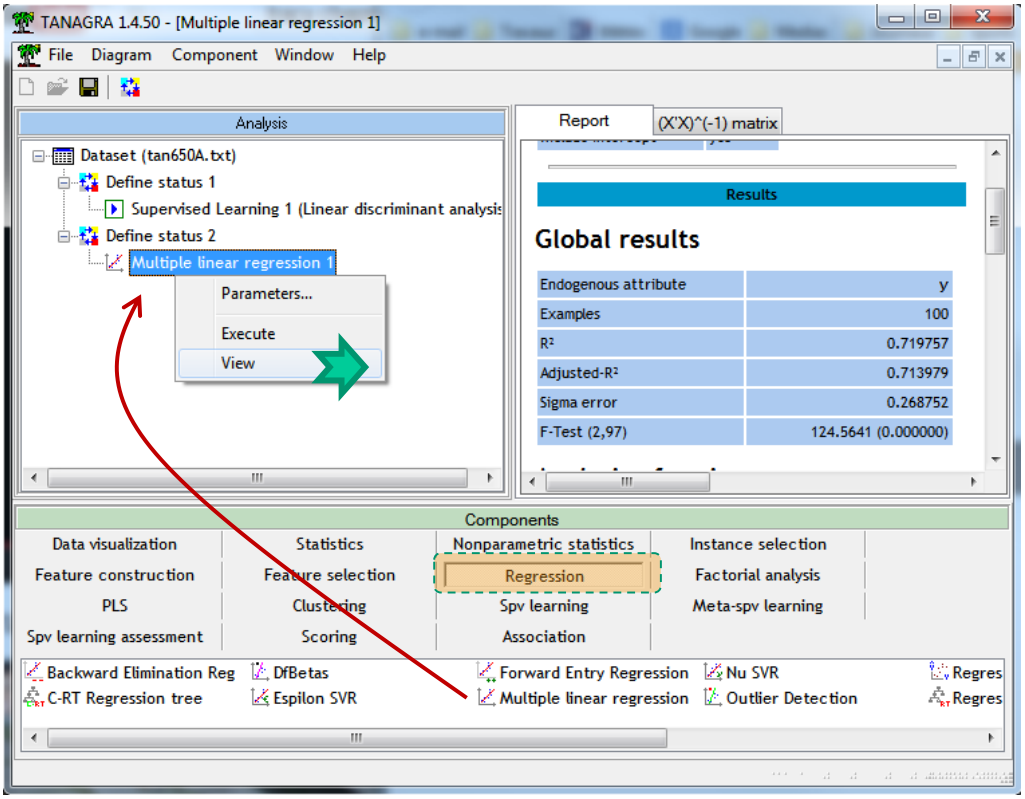
Nous choisissons le codage $\left(y_1 = \frac{n_2}{n} = 0.5; y_2 = -\frac{n_1}{n} = -0.5 \right)$ pour former la variable **y**. Elle correspond à la dernière colonne de notre jeu de données (Figure 1).

2.4.2 Régression linéaire multiple dans Tanagra

Revenons dans Tanagra. Nous insérons un second composant DEFINE STATUS dans le diagramme. Nous plaçons Y en TARGET, PET.LENGTH et PET.WIDTH en INPUT.



Nous insérons ensuite l'outil MULTIPLE LINEAR REGRESSION (onglet REGRESSION). Nous actionnons le menu contextuel VIEW pour consulter les résultats.



Voyons-en le détail.

2.4.3 Evaluation globale du modèle

Le critère R^2 est le principal instrument de l'évaluation globale du modèle. Il indique la proportion de variance expliquée par le modèle. Dans notre cas, $R^2 = 0.719757$.

Global results

Endogenous attribute	y
Examples	100
R²	**0.719757**
Adjusted-R	0.713979
Sigma error	0.268752
F-Test (2,97)	**124.5641 (0.000000)**

Figure 9 - Evaluation globale de la régression

Au R^2, nous pouvons associer la statistique F qui sert à tester la significativité globale du modèle (H0 : les coefficients associés aux variables sont tous nuls, c.-à-d. aucune variable ne contribue à l'explication de Y). Sous H0, elle suit une loi de Fisher à (p, n – p – 1) degrés de liberté. Nous avons F = 124.5641 ; le modèle est globalement significatif à 5% (Figure 9).

2.4.4 Coefficients de la régression et tests de significativité

Le tableau des coefficients fournit les estimations de β_j (**Coef.**) La colonne « **t(97)** » correspond à la statistique du test t_j de significativité (H0 : $\beta_j = 0$). Elle suit une loi de Student à (n – p – 1) degrés de liberté.

Coefficients

Attribute	Coef.	std	t(97)	p-value
pet.length	-0.197641	0.057648	-3.428428	0.000893
pet.width	-0.663436	0.112044	-5.921231	0.000000
Intercept	2.081544	0.168871	12.326226	0.000000

Figure 10 - Coefficients estimés de la régression - Tests de significativité des coefficients

Une rapide comparaison permet d'observer que le rapport entre les coefficients de la fonction score issue de l'analyse discriminante linéaire (LDA) et de l'équation de régression (REG) est le même quelle que soit la variable considérée, y compris la constante (les très légères différences sont dues à des erreurs de troncatures lors de la recopie manuelle des valeurs). Ce phénomène avait été constaté par ailleurs sur une autre fraction du fichier IRIS [setosa vs. versicolor] (Desbois, 2003 ; page 31).

	Score function LDA	Coefficients REG	Ratio
pet.length	-2.764569	-0.197641	13.98783
pet.width	-9.280052	-0.663436	13.98786
constant	29.116340	2.081544	13.98786

Figure 11 - Ratio entre les coefficients de la fonction score (LDA) et de la régression (REG)

De fait, la **régression** pour le classement telle que nous le définissons dans cette section produit un résultat totalement équivalent à celui de **l'analyse discriminante**. Les deux approches construisent **exactement la même frontière** pour séparer les classes.

2.5 Formules de transition et équivalences

L'observer, c'est très bien. Mais le véritable enjeu est de pouvoir calculer ce ratio a priori, afin de déduire les résultats de l'analyse discriminante linéaire (LDA) à partir de la régression multiple linéaire (REG). C'est ce que nous montrons dans cette section.

2.5.1 Du R^2 vers le Λ - Equivalence entre les statistiques d'évaluation globale

Le R^2 de la régression est le rapport entre la variance expliquée par le modèle et la variance totale de la variable cible. Le Λ de Wilks de l'analyse discriminante est le rapport entre les

variances généralisées intra-classes (résiduelle, non expliquée par l'appartenance aux groupes) et totales. Il vient immédiatement la relation suivante :

$$\Lambda = 1 - R^2 = 1 - 0.719757 = 0.280243$$

Nous retrouvons le résultat de la LDA. Les tests de significativités globales sont équivalentes avec F = 124.5641 qui suit une loi de Fisher à (2, 97) degrés de liberté (Figure 4 et Figure 9).

2.5.2 Transition entre les coefficients

Puisque nous avons Λ, nous pouvons calculer la distance de Mahalanobis D^2 entre les centres de classes. Nous obtenons $D^2 = 10.0678$ (voir page 6).

Pour simplifier les notations, posons :

$$c_1 = n_1 + n_2 - 2 = n - 2 = 100 - 2 = 98$$

Et

$$c_2 = \frac{n_1 \times n_2}{n_1 + n_2} = \frac{50 \times 50}{50 + 50} = 25$$

Le ratio entre les coefficients de la LDA et REG s'écrit (Tomassone et al., 1988) :

$$\rho = \frac{\theta_j}{\beta_j} = \frac{c_1 + c_2 \times D^2}{c_2 \times (y_1 - y_2)} \quad (j = 0, 1, \ldots, p)$$

Soit, pour le fichier IRIS :

$$\rho = \frac{98 + 25 \times 10.0678}{25 \times (0.5 - (-0.5))} = 13.98786$$

C'est la valeur trouvée en faisant le rapport entre les coefficients de la LDA et REG (Figure 11). Il est ainsi possible de calculer le coefficient de la 2nde variable (PET.WIDTH) de la fonction score à partir de son homologue de l'équation de régression (Figure 10 et Figure 6) :

$$\theta_2 = \rho \times \beta_2 = 13.98786 \times (-0.663436) = -9.280052$$

Ce rapport ρ est le même quelle que soit le coefficient, y compris la constante lorsque nous sommes dans le cadre des classes équilibrées ($n_1 = n_2$).

2.5.3 Tests de significativité des coefficients

Pour tester la significativité individuelle des coefficients de la régression, nous disposons de la statistique t_j qui suit une loi de Student à (n – p – 1) degrés de liberté. Pour la même finalité dans la discriminante, nous disposons de F_j qui suit une loi de Fisher à (1, n – p – 1) degrés de liberté. Il vient naturellement la relation suivante :

$$F_j = t_j^2$$

Par exemple, pour la première variable explicative (PET.LENGTH) (Figure 10 et Figure 8) :

$$F_1 = t_1^2 = (-3.428428)^2 = 11.75412$$

Ici également, **nous pouvons exploiter directement les résultats de la régression pour expertiser la pertinence des variables dans l'analyse discriminante binaire**.

3 Traitement des classes non équilibrées

La régression produit une constante non proportionnelle à celle de la fonction score lorsque les classes sont déséquilibrées ($n_1 \neq n_2$). La frontière induite par la régression est parallèle à celle de l'analyse discriminante. Par conséquent, le modèle classe différemment puisque la règle de décision est différente (Hastie et al, 2009, page 110). Si l'on souhaite retrouver les résultats de l'analyse discriminante à partir de la régression, il faut introduire une correction supplémentaire pour la constante.

3.1 Correction additionnelle pour la constante

La relation entre les coefficients des variables reste valide c.-à-d. $\theta_j = \rho \times \beta_j$ (j ≥ 1).

Une correction additionnelle δ est nécessaire pour la constante c.-à-d.

$$\tilde{\theta}_0 = \theta_0 + \delta = \rho \times \beta_0 + \delta$$

Cette correction δ dépend de la distribution des classes et des coordonnées des barycentres conditionnels. Elle peut être déduite de l'expression de la fonction score (Nakache et Confais, 2003, page 19 ; les auteurs parlent de fonction linéaire discriminante de Fisher et omettent, de ce fait, la partie relative aux effectifs conditionnels n_1 et n_2) :

$$\delta = \ln \frac{n_1}{n_2} - \frac{1}{2} \sum_{j=1}^{p} \theta_j \times \left[\left(\mu_1^j + \mu_2^j \right) - 2 \times \mu^j \right]$$

Où μ^j correspond à la moyenne de la variable X_j sur l'ensemble des observations, μ_1^j (resp. μ_2^j) la moyenne de la variable X_j pour les observations du groupe G_1 (resp. G_2).

Remarque : On montre facilement que **$\delta = 0$** quand les effectifs sont équilibrés (**$n_1 = n_2$**). En effet, dans cette configuration :

$$\ln\left(\frac{n_1}{n_2}\right) = \ln(1) = 0$$

Et

$$\mu^j = \frac{n_1 \times \mu_1^j + n_2 \times \mu_2^j}{n} = \frac{1}{2}\left(\mu_1^j + \mu_2^j\right) \Rightarrow \left(\mu_1^j + \mu_2^j\right) - 2\mu^j = 0$$

3.2 Données BREAST

Pour illustrer le calcul de la correction pour classes non équilibrées, nous utilisons une fraction du fichier « breast-cancer-wisconsin »[7] limité à **p = 3** variables explicatives (clump, ucellsize, ucellshape). La variable cible TARGET est binaire[8] (**K = 2**), le premier groupe G_1 correspond à la modalité « begnin », G_2 à « malignant ». Nous disposons de **n = 699** observations, avec $\mathbf{n_1 = 458}$ et $\mathbf{n_2 = 241}$. Voici les premières lignes du jeu de données.

clump	ucellsize	ucellshape	target
4	2	2	begnin
1	1	1	begnin
2	1	1	begnin
10	6	6	malignant
4	1	1	begnin

3.2.1 Codage de la variable cible

La première étape consiste à recoder la variable cible, nous créons Y qui prend deux valeurs possibles : $y_1 = \frac{n_2}{n} = \frac{241}{699} = 0.345$ et $y_2 = -\frac{n_1}{n} = -\frac{458}{699} = -0.655$

3.2.2 Résultats de la régression

Nous envoyons Y et les p = 3 variables prédictives dans Tanagra, nous réalisons la régression, nous obtenons les résultats suivants.

Global results

Endogenous attribute	y
Examples	699
R^2	0.747486
Adjusted-R^2	0.746396
Sigma error	0.239526
F-Test (3,695)	685.7753 (0.000000)

Coefficients

Attribute	Coef.	std	t(695)	p-value
clump	-0.048006	0.004315	-11.124401	0
ucellsize	-0.053245	0.007144	-7.453079	0
ucellshape	-0.051713	0.007415	-6.973756	0
Intercept	0.544840	0.016888	32.262169	0

[7] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

[8] Nous avons modifié le nom de la variable « class » en « target » pour éviter les confusions lors du traitement sous R.

A ce stade, nous disposons de tous les éléments pour calculer le rapport ρ entre les coefficients de la fonction score et de la régression.

3.2.3 Calcul du ratio ρ - Obtention de la fonction score

Plusieurs étapes sont nécessaires pour y parvenir. Nous devons tout d'abord déduire le Λ de Wilks à partir du R² :

$$\Lambda = 1 - R^2 = 1 - 0.747486 = 0.252514$$

Puis nous calculons la distance de Mahalanobis D² :

$$D^2 = \frac{1-\Lambda}{\Lambda} \times \frac{n(n-2)}{n_1 \times n_2} = \frac{1-0.252514}{0.252514} \times \frac{699(699-2)}{458 \times 241} = 13.06607$$

Calculons c_1 et c_2 pour être cohérent avec la présentation de la section précédente :

$$c_1 = n_1 + n_2 - 2 = n - 2 = 699 - 2 = 697$$

$$c_2 = \frac{n_1 \times n_2}{n_1 + n_2} = \frac{458 \times 241}{458 + 241} = 157.908$$

Nous obtenons enfin ρ

$$\rho = \frac{c_1 + c_2 \times D^2}{c_2 \times (y_1 - y_2)} = \frac{697 + 157.908 \times 13.06607}{157.908 \times (0.345 - (-0.655))} = 17.48002$$

A partir des coefficients de la régression β_j, nous pouvons calculer les coefficients de la fonction score $\theta_j = \beta_j \times \rho$:

	Beta_j	Theta_j
clump	-0.048006	-0.83915
ucellsize	-0.053245	-0.93072
ucellshape	-0.051713	-0.90394
Intercept	0.544840	9.52382

3.2.4 Correction de la constante δ

Les coefficients associés aux variables sont ceux de la LDA. Il faut corriger la constante en revanche. Nous devons tout d'abord calculer les barycentres globaux et conditionnels pour ce faire.

	Barycentres		
Classes	_clump	_ucellsize	_ucellshape
mu_1	2.956	1.325	1.443
mu_2	7.195	6.573	6.560
mu	**4.418**	**3.134**	**3.207**

Pour obtenir δ :

$$\delta = \ln\frac{458}{241} - \frac{1}{2}\{-0.83915\times[(2.956+7.195)-2\times 4.418]+\cdots\} = 2.67021$$

Ainsi, la constante corrigée s'écrit :

$$\tilde{\theta}_0 = \theta_0 + \delta = 9.52382 + 2.67021 = 12.19403$$

La fonction score obtenue par ce truchement s'écrit finalement :

	Score function LDA by REG
clump	**-0.83915**
ucellsize	**-0.93072**
ucellshape	**-0.90394**
Intercept	**12.19403**

3.2.5 Confrontation avec la fonction score issue de l'analyse discriminante

Voyons ce qu'il en est lorsque nous réalisons directement une analyse discriminante sous Tanagra.

MANOVA

Stat	Value	p-value
Wilks' Lambda	0.2525	-
Bartlett -- C(3)	957.2095	0
Rao -- F(3, 695)	685.7753	0

LDA Summary

	Classification		Score	Statistical Evaluation			
Attribute	begnin	malignant	Function	Wilks L.	Partial L.	F(1,695)	p-value
clump	0.70839	1.54754	**-0.83915**	0.297477	0.848853	123.75231	0
ucellsize	0.13147	1.06218	**-0.93072**	0.272696	0.92599	55.54839	0
ucellshape	0.25922	1.16318	**-0.90395**	0.270184	0.9346	48.63328	0
constant	-1.74408	-13.93812	**12.19404**	-			

Figure 12 - Résultats de la LDA - Données "breast"

Nous retrouvons bien les mêmes valeurs. Les petites différences sont dues aux erreurs de troncature dans les calculs intermédiaires.

3.3 Un exemple de traitement sous R

Pour que le lecteur puisse facilement reproduire la démarche et, pourquoi pas, la transposer à d'autres fichiers, je propose de reprendre tout le processus sous forme de programme R dans cette section. Voici le code source commenté.

```
#importation des données
library(xlsx)
breast <- read.xlsx(file="lda_regression_dataset.xls", header=T, sheetIndex=2)
print(summary(breast))

#effectifs
n1 <- table(breast$target)[1] #begnin
n2 <- table(breast$target)[2] #malignant
n <- n1+n2

#codage variable cible  - Tomassone, page 38
y1 <- n2/n
y2 <- -n1/n
y <- ifelse(breast$target=="begnin",y1,y2)

#régression sur variable recodée
reg <- lm(y ~ ., data = breast[-4])
print(reg)
beta <- reg$coefficients
print(round(beta,5))

#summary de la régression
sreg <- summary(reg)

#R2 de la régression
R2 <- sreg$r.squared

#D2 (distance de Mahalanobis) - Huberty, page 353 ; Tomassone, page 38
D2 <- (R2/(1-R2))*(n*(n-2))/(n1*n2)
names(D2)[1] <- "D2"
print(D2)

#quantités intermédiaires pour calculs (Tomassone, page 27)
c1 <- n1+n2-2
c2 <- (n1*n2)/(n1+n2)
```

```
#rho – facteur de correction
rho <- (c1+c2*D2)/(c2*(y1-y2))
print(rho)

#fonction score avant correction de la constante
theta <- beta*rho
print(round(theta,5))

#*** correction de la constante ***

#tenir compte de la prévalence – 1ere correction
e1 <- log(n1/n2)

#moyennes
mu <- sapply(breast[1:3],mean)

#moyennes conditionnelles
mu.cond <- aggregate(breast[1:3],by=list(breast$target),mean)[2:4]

#moyennes conditionnelles sur variables centrées
mu.centre <- ((mu.cond[1,]+mu.cond[2,])-2*mu)

#coef. de la LDA sur variables (sans la constante)
coef.lda.p <- theta[2:4]

#produit scalaire - 2nde correction
e2 <- -0.5*sum(coef.lda.p*mu.centre)

#delta
delta <- e1 + e2
print(delta)

#correction de la constante de la LDA
theta_tilde <- theta
theta_tilde[1] <- theta[1] + delta

#fonction score LDA après correction de la constante
print(round(theta_tilde,5))

#*** comparaison des performances ***
```

```
#fonction pour matrice de confusion et taux d'erreur
confusion.matrix <- function(dataset,coef){
 #prediction for one row
 prediction <- function(data.row){
  score <- sum(data.row[1:3]*coef[2:4])+coef[1]
  return(ifelse(score>=0,"begnin","malignant"))
 }
 #prediction for all rows
 pred <- factor(apply(data.matrix(dataset),1,prediction))
 #confusion matrix
 cm <- table(dataset$target,pred)
 print(cm)
 #error rate
 er <- 1-sum(diag(cm))/sum(cm)
 print(er)
}

#matrice de confusion regression
confusion.matrix(breast,beta)

#matrice de confusion lda
confusion.matrix(breast,theta_tilde)
```

Voici les principales sorties du programme.

Coefficients de la régression β.

```
> print(round(beta,5))
(Intercept)       clump   ucellsize  ucellshape
    0.54484    -0.04801    -0.05324    -0.05171
```

Distance de Mahalanobis D^2 obtenue à partir du R^2 de la régression.

```
> print(D2)
      D2
13.06609
```

Calcul du ratio ρ.

```
> print(rho)
     rho
17.48004
```

1ère version de la fonction score.

```
> print(round(theta,5))
(Intercept)       clump   ucellsize  ucellshape
    9.52382    -0.83915    -0.93072    -0.90395
```

Correction de la constante δ.

```
> print(delta)
   delta
2.670214
```

Fonction score après correction de la constante (voir Figure 12).

```
> print(round(theta_tilde,5))
(Intercept)       clump   ucellsize  ucellshape
   12.19404    -0.83915    -0.93072    -0.90395
```

Comparaison des performances.

```
> #matrice de confusion regression
> confusion.matrix(breast,beta)
           pred
            begnin malignant
  begnin       435        23
  malignant      9       232
[1] 0.04577969
>
> #matrice de confusion lda
> confusion.matrix(breast,theta_tilde)
           pred
            begnin malignant
  begnin       448        10
  malignant     33       208
[1] 0.06151645
```

Très curieusement, la régression (taux d'erreur = 4.58%) serait finalement plus performante que l'analyse discriminante (taux d'erreur = 6.15%) sur notre fichier. Il ne faut surtout pas en tirer des conclusions hâtives. Il s'agit d'un exemple sur un seul fichier, et les performances sont évaluées en resubstitution c.-à-d. nous utilisons les mêmes données pour l'apprentissage et le test. Notons simplement que la régression (avant correction de la constante) et l'analyse discriminante produisent bien des modèles différents lorsque les classes sont déséquilibrées. C'est ce qui explique les disparités entre les matrices de confusion.

4 Conclusion

Travailler sur ce tutoriel a été particulièrement passionnant. Je savais depuis longtemps qu'il était possible d'obtenir les résultats de l'analyse discriminante binaire à partir de la régression linéaire multiple puisque tout le monde en parlait. Mais c'est tout autre chose que d'aller dans le détail lorsqu'il faut expliciter la démarche dans un tutoriel. Il fallait identifier la bonne formule de transition à chaque étape, et reconstruire l'expression permettant de rectifier la constante lorsque les classes ne sont pas équilibrées. Il est étonnant que cette correction ne soit pas plus abondamment décrite dans les articles et ouvrages qui font référence.

La synonymie entre ces deux approches dépasse la simple curiosité scientifique. Elle rend licite l'utilisation d'un programme / algorithme de régression multiple pour réaliser une l'analyse discriminante linéaire binaire. Les résultats sont totalement équivalents, mais les traitements seront plus rapides parce que les calculs et les structures de données sont plus simples pour la régression, augmentant d'autant notre capacité à traiter les grandes bases de données (volumétrie, temps de traitements). Cet avantage prend une acuité singulière dans un processus de sélection automatique de variables où les calculs sont particulièrement ardus, fortement consommateurs de ressources machines.

5 Bibliographie

C.M. Bishop, « Pattern Recognition and Machine Learning », Springer, 2007.

D. Desbois, « Une introduction à l'analyse discriminante avec SPSS pour Windows », Revue Modulad, n°30, 2003.

R.O. Duda, P.E. Hart, D. Stork, « Pattern Classification », 2nd Edition, Wiley, 2000.

T. Hastie, R. Tibshirani, J. Friedman, « The Elements of Statistical Learning », Springer, 2009.

C.J. Huberty, S. Olejnik, « Applied MANOVA and Discriminant Analysis »,Wiley, 2006.

J.P. Nakache, J. Confais, « Statistique explicative appliquée », Technip, 2003.

G. Saporta, « Probabilités, Analyse des Données et Statistique », Technip, 2006.

R. Tomassone, M. Danzart, J.J. Daudin, J.P. Masson, « Discrimination et Classement », Masson, 1988.