Régression linéaire multiple pour le classement

Comment la transposer dans un problème de classement ?

Ricco Rakotomalala
Université Lumière Lyon 2

Utiliser la régression dans un problème de classement binaire

(Y à K = 2 modalités)

Analyse prédictive : régression vs. classement

Régression

- Y continue à prédire
- X prédictives, quelconques

Classement

- Y discrète à prédire
- X prédictives, quelconques

On veut construire une fonction de prédiction (explication) telle que

$$Y = f(X, \alpha)$$

Problèmes :
- il faut choisir une famille de fonction
- il faut estimer les paramètres α
- on utilise un échantillon pour optimiser sur la population

Critères d'évaluation

Erreur quadratique
Somme des carrés des erreurs

$$S = \sum_{\Omega} [Y - \hat{f}(X, \hat{\alpha})]^2$$

Taux d'erreur
Erreur 0/1 (bon ou mauvais classement)

$$ET = \frac{1}{card(\Omega)} \sum_{\Omega} \Delta[Y, \hat{f}(X, \hat{\alpha})]$$

$$où\ \Delta[.] = \begin{cases} 1 si Y \neq \hat{f}(X, \hat{\alpha}) \\ 0 si Y = \hat{f}(X, \hat{\alpha}) \end{cases}$$

Régression linéaire multiple : rappel

- Se restreindre à une famille de fonction de prédiction linéaire
- Et à des exogènes continues (éventuellement des qualitatives recodées)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i \; ; \; i = 1, \ldots, n$$

Le terme aléatoire ε - l'erreur du modèle – cristallise les « insuffisances » du modèle :

- le modèle n 'est qu'une caricature de la réalité, la spécification (linéaire notamment) n 'est pas toujours rigoureusement exacte
- des variables qui ne sont pas prises en compte dans le modèle
- les fluctuations liées à l 'échantillonnage (si on change d 'échantillon, on peut obtenir un résultat différent)

$\hat{\varepsilon}$ quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle

$(a_0, a_1, \ldots, a_p)$ sont les paramètres du modèle que 'l'on veut estimer à l'aide des données

Régression sur variable cible recodée : le cas binaire -- Y∈{+, -}

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} 1, \text{si } y_i = + \\ 0, \text{si } y_i = - \end{cases}$$

On constate aisément

$$E(Z_i) = P(Y_i = +)$$

Reportée dans l'équation de régression

$$E(Z_i) = P(Y_i = +) = a_0 + a_1 x_{i,1} + \ldots + a_p x_{i,p}$$

On devrait donc pouvoir mettre en place une régression qui permet d'estimer directement la probabilité d'appartenance P(Y=+) ???

hélas, non...

» la combinaison linéaire varie entre $-\infty$ et $+\infty$, ce n'est pas une probabilité
» les hypothèses de la MCO, notamment l'homoscédasticité et la normalité de l'erreur posent problèmes

Régression : point de vue géométrique

Vue probabiliste

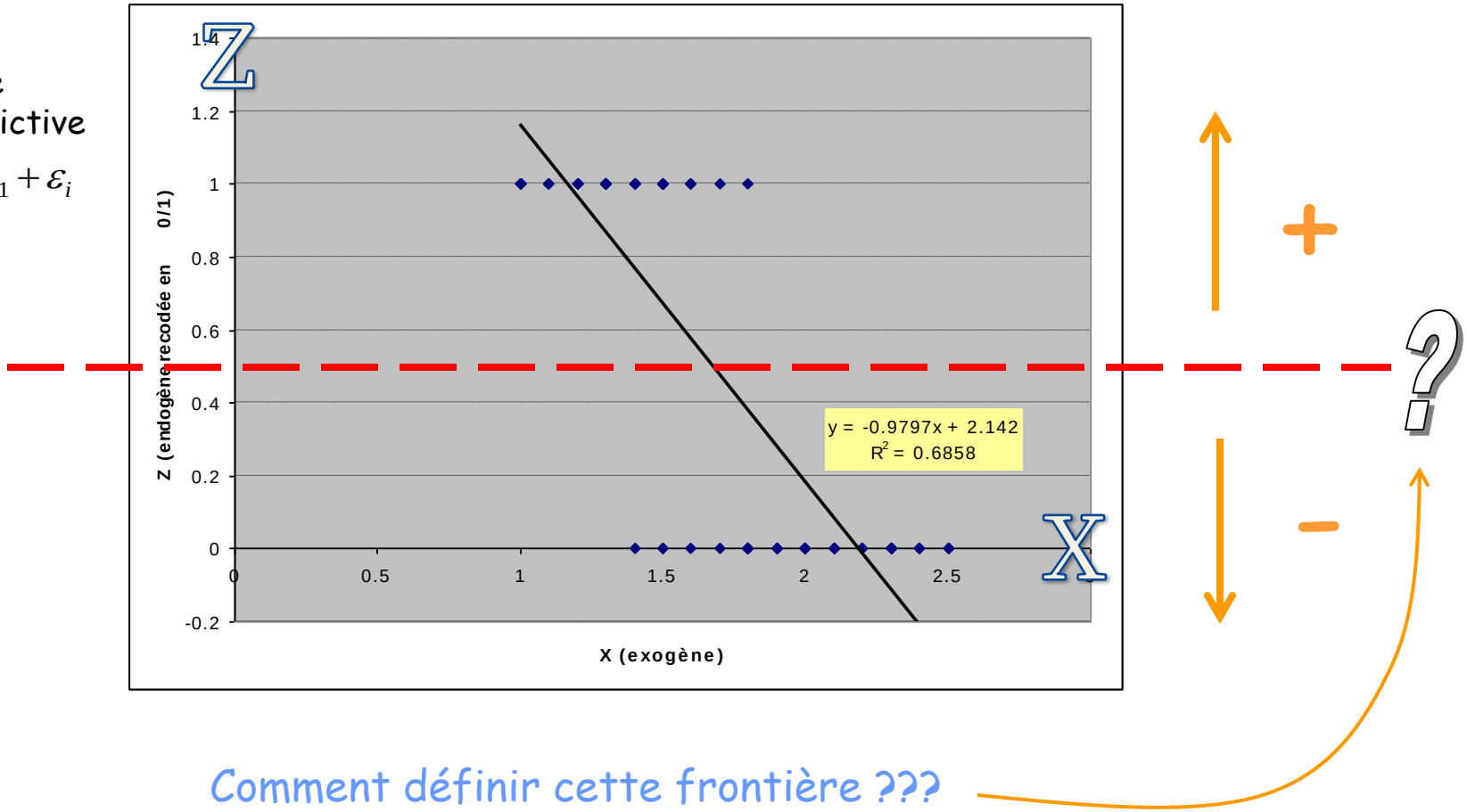
La régression linéaire n'est pas adéquate pour estimer P(Y=+/X), …

Mais on peut exploiter ses résultats pour « séparer » des groupes !!!

Vue géométrique

Ex. Une seule variable prédictive

$$z_i = a_0 + a_1 x_{i,1} + \varepsilon_i$$



Comment définir cette frontière ???

Régression : règle d'affectation avec codage 0/1

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} 1, \text{si } y_i = + \\ 0, \text{si } y_i = - \end{cases}$$

On effectue la régression linéaire multiple (DROITEREG dans EXCEL par ex.)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

EMCO

$$\hat{z}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \hat{a}_2 x_{i,2} + \cdots + \hat{a}_p x_{i,p}$$

Règle d'affectation

$$\hat{y}_i = \begin{cases} +, \text{si } \hat{z}_i > \bar{z} \\ -, \text{si } \hat{z}_i \leq \bar{z} \end{cases}$$

Moyenne des « z » c.-à-d. $\bar{z} \approx P(Y = +)$

Régression : règle d'affectation avec autre codage

Dans un cas Y binaire (Positifs vs. Négatifs), nous pouvons coder

$$z_i = \begin{cases} \frac{n_-}{n}, \text{si } y_i = + \\ -\frac{n_+}{n}, \text{si } y_i = - \end{cases}$$

On effectue la régression linéaire multiple (DROITEREG dans EXCEL par ex.)

$$z_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i$$

EMCO

$$\hat{z}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \hat{a}_2 x_{i,2} + \cdots + \hat{a}_p x_{i,p}$$

Règle d'affectation

$$\hat{y}_i = \begin{cases} +, \text{si } \hat{z}_i > 0 \\ -, \text{si } \hat{z}_i \leq 0 \end{cases}$$

On remarque ici

$$\bar{z} = \frac{1}{n}\left(n_+ \times \frac{n_-}{n} + n_- \times (-\frac{n_+}{n}) \right) = 0$$

Equivalence avec l'analyse discriminante

(Y à K = 2 modalités)

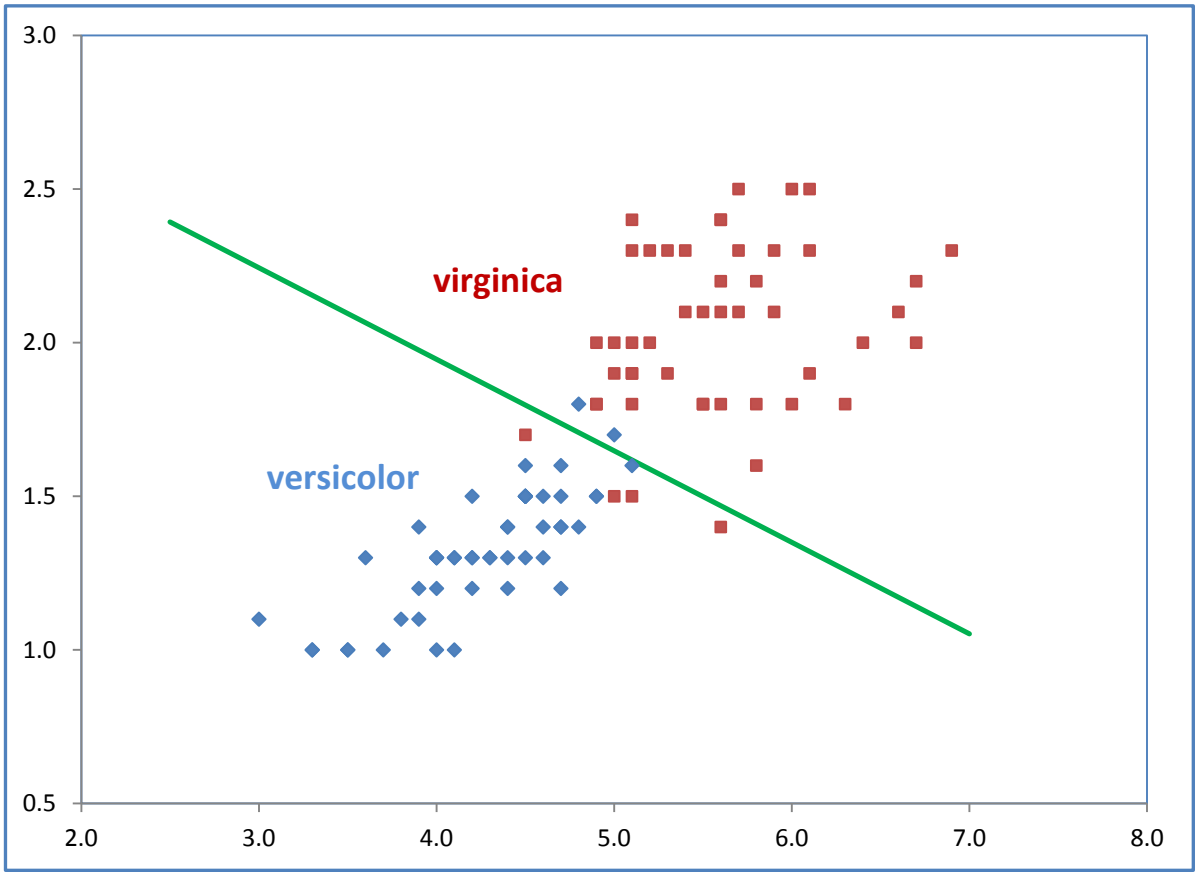
Discrimination : construire une frontière linéaire

n = 100 individus

Dans le plan c.-à-d. p = 2 variables prédictives.

Cible à K = 2 modalités c.-à-d. 2 groupes d'individus ($n_1 = n_2 = 50$)

Les méthodes supervisées linéaires visent à produire cette frontière.



Comparaisons des résultats

Régression

Analyse discriminante

Global results

R^2	0.7198
Adjusted-R	0.713979
Sigma error	0.268752
F-Test (2,97)	124.5641 (0.000000)

Coefficients

Attribute	Coef.	std	t(97)	p-value
pet.length	-0.198	0.057648	-3.428	0.000893
pet.width	-0.663	0.112044	-5.921	0.000000
Intercept	2.082	0.168871	12.326	0.000000

L'équivalence est totale !

MANOVA

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

$$\Lambda = 1 - R^2 = 1 - 0.7198 = 0.2802$$

$$F_j = t_j^2$$

$$11.754 = (-3.428)^2, \ldots$$

LDA Summary

	Classification functions		Score function	Statistical Evaluation			
Attribute	versicolor	virginica	D(X)	Wilks L.	Partial L.	F(1,97)	p-value
pet.length	14.40029	17.164859	-2.765	0.314202	0.89192	11.754	0.000893
pet.width	7.824622	17.104674	-9.280	0.381538	0.734509	35.061	0.000000
constant	-36.55349	-65.66983	29.116	-			

$$\theta_j = \beta_j \times \rho$$

$$-2.765 = -0.198 \times 13.988$$

$$-9.280 = -0.663 \times 13.988$$

$$29.116 = 2.082 \times 13.988$$

On sait calculer directement ρ !

Cas des classes non équilibrées ($n_1 \neq n_2$)

$n = 183$ avec
$n_1 = 96$, $n_2 = 87$

Régression

Global results

R^2	0.2753
Adjusted-R	0.2672
Sigma error	0.4287
F-Test (2,180)	**34.1851**

Coefficients

Attribute	Coef.	std	t(180)	p-value
max.rate	-0.0076	0.0014	-5.3940	0.0000
oldpeak	0.1701	0.0327	5.1990	0.0000
Intercept	0.8463	0.2200	3.8461	0.0002

Analyse discriminante

MANOVA

Stat	Value	p-value
Wilks' Lambda	0.7247	-
Bartlett -- C(2)	57.9534	0
Rao -- F(2, 180)	**34.1851**	0

$$\Lambda = 1 - R^2 = 1 - 0.2753 = 0.7247$$

$(-5.3940)^2 = 29.0951$
$(5.1990)^2 = 27.0301$

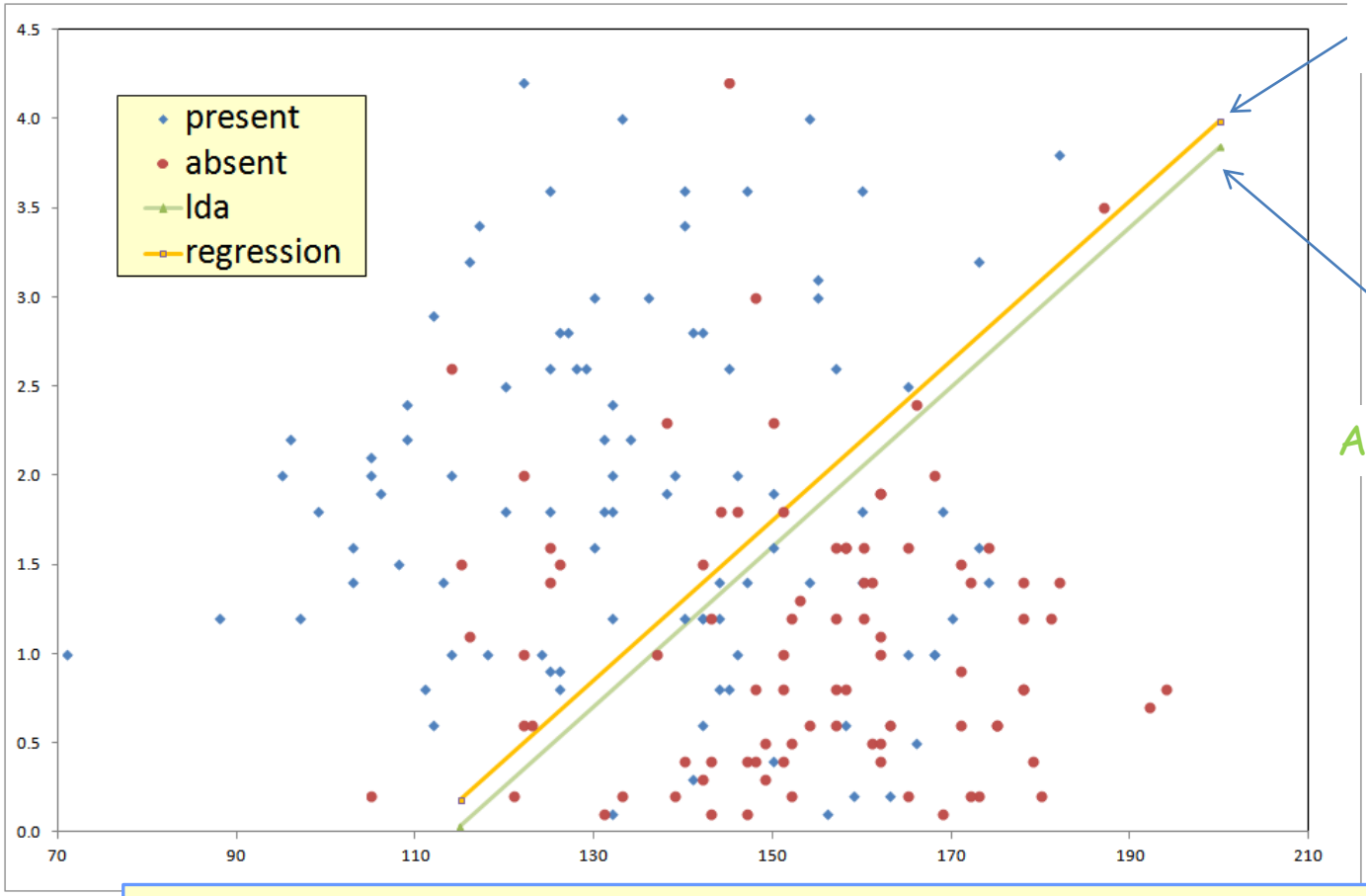
LDA Summary

	Classification functions		Fonction score	Statistical Evaluation			
Attribute	present	absent		Wilks L.	Partial L.	F(1,180)	p-value
max.rate	0.3113	0.3530	-0.0417	0.8419	0.8609	29.0951	0.0000
oldpeak	2.3975	1.4665	0.9310	0.8336	0.8694	27.0301	0.0000
constant	-23.9246	-28.6913	4.7667	-			

$-0.0417 / -0.0076 = 5.4721$
$0.9310 / 0.1701 = 5.4721$
$4.7667 / 0.8463 = 5.6323$

Il n'y a plus l'équivalence pour la constante. La règle de décision est différente !!!

Cas des classes non équilibrées – Frontières induites



Régression linéaire

Analyse discriminante

(1) Seule la constante diffère
(2) Les frontières sont différentes mais parallèles
(3) Les modèles se comportent différemment c.-à-d. les matrices de confusion ne sont pas identiques
(4) L'amplitude de l'écart dépend de l'amplitude du déséquilibre entre les classes

Régression vs. Analyse discriminante linéaire - BILAN

Il est possible de retrouver les coefficients de l'ADL à partir d'une régression linéaire multiple
>> l'équivalence est totale dans le cas des classes équilibrées
>> la constante est différente lorsque $n_1 \neq n_2$, il faudrait une correction additionnelle

Attention, les hypothèses ne sont pas les mêmes, notamment :
- les X sont aléatoires dans la discrimination
- Y est binaire (dist. binomiale) et non pas normale (dist. normale) par l'intermédiaire du terme d'erreur ε

Il n'en reste pas moins que les tests de significativité globale du modèle et individuelle des coefficients sont directement utilisables, quelle que soit la distribution des classes

Comparaison des méthodes

Régression pour le classement

Analyse discriminante linéaire

Régression logistique

Trois séparateurs linéaires (Rég. Logistique, Analyse Discriminante, Rég. Linéaire)

Régression logistique

$$LOGIT = \ln \frac{P(Y=+/X)}{1-P(Y=+/X)} = \ln \frac{P(Y=+/X)}{P(Y=-/X)} = a_0 + a_1 x_1 + \cdots + a_p x_p$$

$$\hat{Y} = + \; si \; LOGIT > 0$$

Analyse discriminante linéaire

$$D(X) = \{\ln[P(Y=+) \times P(X/Y=+)]\} - \{\ln[P(Y=-) \times P(X/Y=-)]\}$$
$$= b_0 + b_1 x_1 + \cdots + b_p x_p$$

$$\hat{Y} = + \; si \; D(X) > 0$$

Régression linéaire multiple

$$Z = c_0 + c_1 x_1 + \cdots + c_p x_p \; ; \; Z = \begin{cases} 1, Y = + \\ 0, Y = - \end{cases}$$

$$\hat{Y} = + \; si \; \hat{Z} > \overline{Z}$$

Fichier BREAST (Y binaire « cancer », 9 descripteurs continus) – Évaluation en resubstitution

Régression logistique

Attribute	Coef.	Std-dev	Wald	Signif
clump	-0.531	0.132	16.237	0.000
ucellsize	-0.006	0.187	0.001	0.975
ucellshape	-0.333	0.208	2.567	0.109
mgadhesion	-0.240	0.115	4.380	0.036
sepics	-0.069	0.151	0.212	0.645
bnuclei	-0.400	0.089	20.041	0.000
bchromatin	-0.411	0.156	6.918	0.009
normnucl	-0.145	0.102	2.003	0.157
mitoses	-0.551	0.303	3.311	0.069
constant	9.671	-	-	-

	begnin	malignant	Sum
begnin	447	11	458
malignant	11	230	241
Sum	458	241	699

$$\varepsilon = \frac{11+11}{699} = 0.0315$$

Analyse discriminante linéaire

	Classification		Statistical Evaluation			
Attribute	begnin	malignant	Wilks L.	Partial L.	F(1,689)	p-value
clump	0.729	1.616	0.184	0.892	83.767	0.000
ucellsize	-0.316	0.292	0.167	0.983	12.264	0.000
ucellshape	0.066	0.504	0.165	0.990	6.662	0.010
mgadhesion	0.057	0.232	0.164	0.996	2.608	0.107
sepics	0.654	0.870	0.164	0.997	2.290	0.131
bnuclei	0.209	1.427	0.210	0.779	195.186	0.000
bchromatin	0.686	1.245	0.168	0.977	16.553	0.000
normnucl	0.000	0.462	0.169	0.971	20.885	0.000
mitoses	0.201	0.278	0.164	1.000	0.324	0.569
constant	-3.048	-23.296	-			

	begnin	malignant	Sum
begnin	448	10	458
malignant	18	223	241
Sum	466	233	699

$$\varepsilon = \frac{18+10}{699} = 0.0401$$

Régression linéaire multiple (begnin = 1)

Attribute	Coef.	std	t(689)	p-value
clump	-0.033	0.004	-9.152	0.000
ucellsize	-0.023	0.006	-3.502	0.000
ucellshape	-0.016	0.006	-2.581	0.010
mgadhesion	-0.006	0.004	-1.615	0.107
sepics	-0.008	0.005	-1.513	0.131
bnuclei	-0.045	0.003	-13.971	0.000
bchromatin	-0.021	0.005	-4.069	0.000
normnucl	-0.017	0.004	-4.570	0.000
mitoses	-0.003	0.005	-0.569	0.569
Constant	1.253			

	begnin	malignant	Sum
begnin	442	16	458
malignant	4	237	241
Sum	466	233	699

$$\varepsilon = \frac{4+16}{699} = 0.0286$$

Conclusion

Ca marche ! (pour le cas binaire K = 2)

Conclusion

(1) On peut utiliser la régression linéaire pour le classement binaire.

(2) Du point de vue de l'inférence statistique, c'est discutable ; du point de vue géométrique, ça tient la route.

(3) Dans le cas binaire, il y a une équivalence avec l'analyse discriminante.

(4) Elle est totale dans le cas des classes équilibrées ($n_1 = n_2$).

(5) La constante est différente dans le cas contraire ($n_1 \neq n_2$). Mais les coefficients des variables et les tests de significativité restent valables. On peut utiliser un programme de sélection de variables de la régression pour l'analyse discriminante.

(6) Il y a différentes solutions possibles pour ($K > 2$). Mais elles présentent différents inconvénients (les frontières entres les classes peuvent être incohérentes), et il n'y a plus l'équivalence avec l'analyse discriminante.

Bibliographique

• Tutoriel Tanagra, « Analyse discriminante et régression linéaire », Avril 2014.

• R. Tomassone, M. Danzart, J.J. Daudin, J.P. Masson, « Discrimination et classement », Masson, 1988, pages 36 à 38

• G. Saporta, « Probabilités, Analyses de données et Statistique », Technip, 2006, pages 451 et 452.