Linear regression for classification

How to use (can we use) the multiple linear regression method for a classification problem ?

Ricco Rakotomalala
Université Lumière Lyon 2

Using the regression in a binary classification problem

(Target variable Y with K = 2 levels)

Supervised learning : continuous vs. discrete target attribute

Regression analysis

- Y continuous target attribute
- X descriptors, continuous or discrete

Classification problem

- Y discrete target attribute
- X descriptors, cont. or disc.

We want to construct a prediction function **f(.)** such as …

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

Problems:
- choosing the function f(.)
- estimating its parameters α
- all the calculations are based on a sample

Evaluating the quality of the predictions

Quadratic error function
Sum of squared error

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

Error rate
0/1 loss (good or bad classification)

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

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

Multiple linear regression: a reminder

- Modeling with linear prediction function
- Continuous dependent variable Z
- Continuous (or dummy coded) explanatory variables, X_1, X_2, …

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

The error term ε captures all the factors which influence the dependent variable other than the explanatory variables i.e.

- le relationship between the dependent and the explanatory variables is not necessarily linear
- some relevant variables are not included in the model
- sampling fluctuation

$\hat{\varepsilon}$ is the residual, this is the difference between the observed value of the dependent variable and its estimated value by the model

$(a_0, a_1, \ldots, a_p)$ is the parameter vector, we want to estimate its values on a sample

Linear regression of an indicator variable : the binary case -- Y∈{+, -}

In the two classes problem (Positive vs. Negative), we can code the target variable Y as follows

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

We observe that

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

Thus…

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

Can we use the linear regression to estimate the posterior probability P(Y=+ / X) ???

No, because…

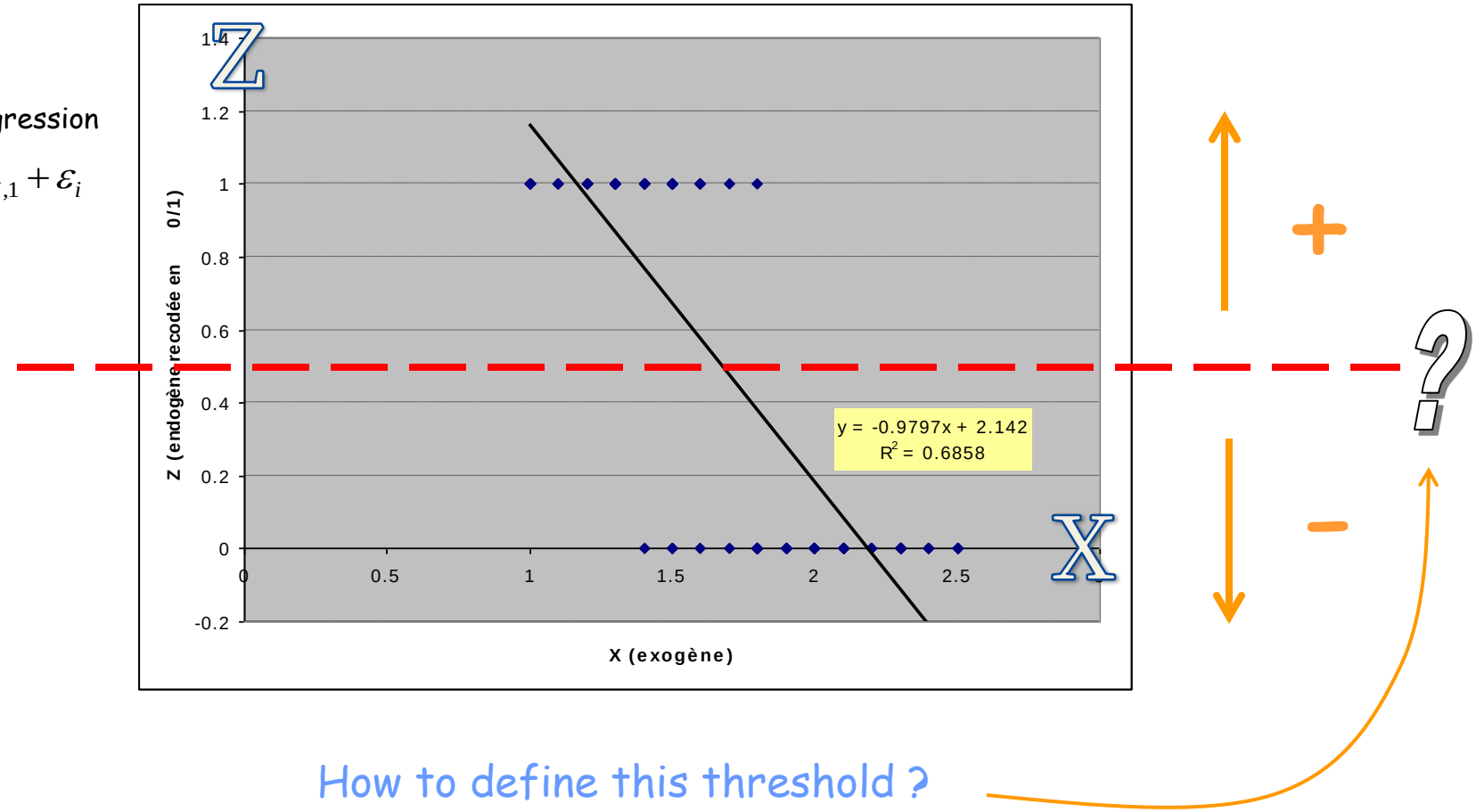
>> the linear combination is defined between -∞ and + ∞, this is not a probability
>> the assumptions under the OLS approach are violated

Simple linear regression: a geometrical point of view

The linear combination cannot used to estimate the probability P(Y=+/X), …
But it can be used to separate the groups !!!

E.g. Linear regression

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



How to define this threshold ?

Decision rule with the 0/1 coding of the target attribute

For a two classes problem, we can code the target attribute as follows:

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

We perform the linear regression (OLS: ordinary least squares method)

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

We obtain the estimated coefficients

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

Decision rule

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

Mean of « z » i.e. $\bar{z} \approx P(Y = +)$

Decision rule with another coding scheme

We can use another coding scheme

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

Regression analysis

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

OLS estimators

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

Decision rule

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

We observe that…

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

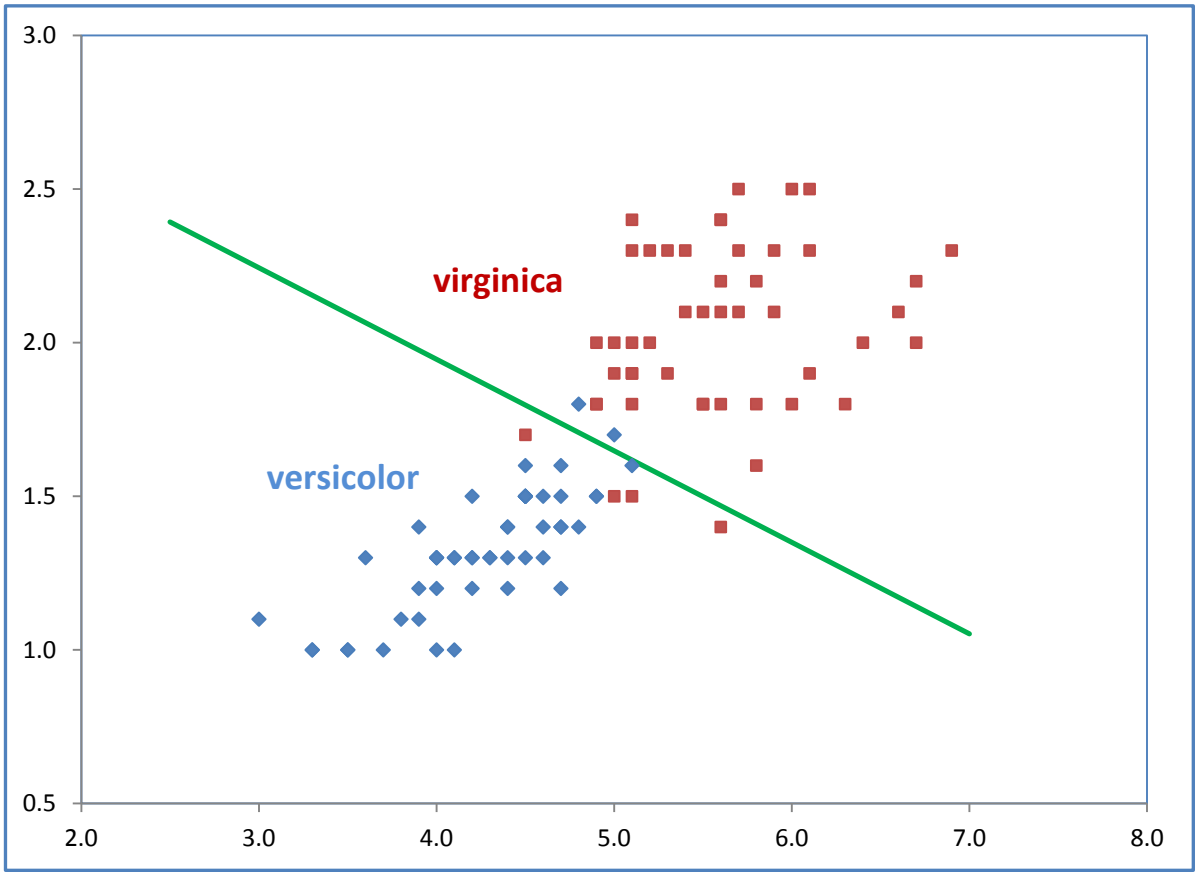
Equivalence between regression and linear discriminant analysis in the binary case

(Y with K = 2 levels)

Linear classifier: a straight line to separate the groups

n = 100 instances
p = 2 predictive variables
K = 2 groups with (n_1 = n_2 = 50)

The linear approaches induces a linear frontier to separate the groups.



Equivalence between the results of regression and linear discriminant analysis

Regression

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

There is exact correspondence!

Discriminant analysis

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	-			

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

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

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

$$29.116 = 2.082 \times 13.988$$

We know how to calculate ρ directly!

When the classes are not balanced ($n_1 \neq n_2$)

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

Regression

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

Discriminant analysis

MANOVA

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

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

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

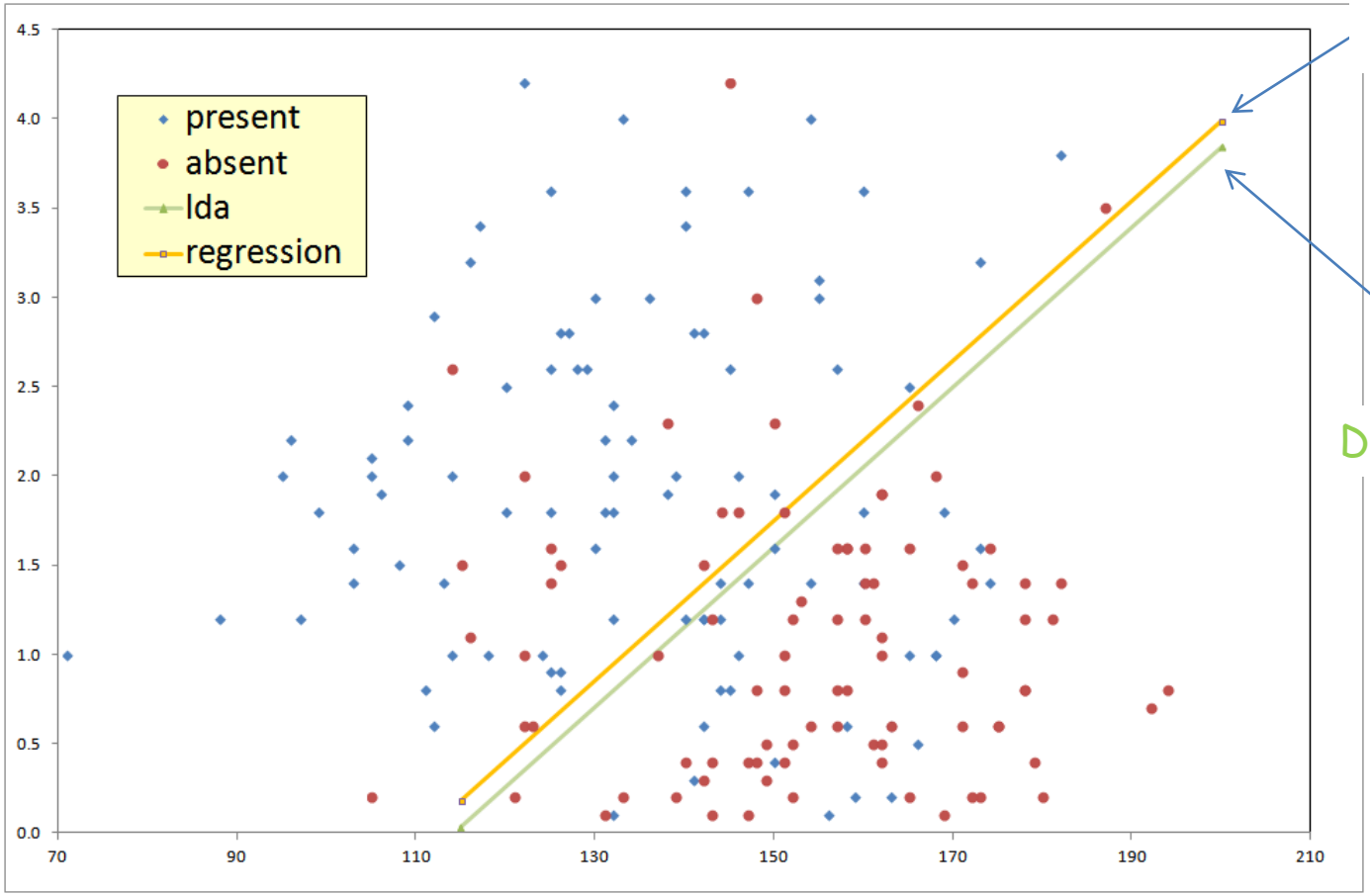
LDA Summary

	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	-			

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

The intercepts are different. The decision rules are different !!!

The induced frontiers when the classes are not balanced



Linear regression

Discriminant analysis

(1) The intercepts are different
(2) We have parallel lines to separate the groups
(3) The model performances are different i.e. the confusion matrices are different
(4) The magnitude of the gap depends on the degree of class imbalance

Regression vs. Linear discriminant analysis - Equivalence

We can obtain the coefficients of the linear discriminant function from the results of the linear regression
>> the models are exactly the same for balanced data
>> the intercepts are different when $n_1 \neq n_2$, an additional correction is needed

Warning, the statistical assumptions under the methods are not identical:
- X are treated as fixed values in regression
- the error term is particular to the regression
- etc.

Nevertheless, we can use the test for global significance of the model and the significance tests for coefficients, whatever the class distribution (balanced or imbalanced case).

Comparison of classifiers

Regression for the classification

Linear discriminant analysis

Logistic regression

Three linear classifiers

Logistic regression

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

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

Linear discriminant analysis

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

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

Multiple linear regression for classification

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

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

BREAST CANCER dataset (Binary target, 9 descriptors) – Resubstitution error rate

Logistic regression

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

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

Linear discriminant analysis

	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

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

Linear regression (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

We can use the regression for a binary classification problem (K = 2)

Conclusion

(1) We can use the linear regression for a binary classification problem.

(2) From the statistical point of view, it may be questionable; from the geometrical point of view, we can find some justifications.

(3) In the binary case (**K = 2**), the regression is equivalent to the discriminant analysis.

(4) All the coefficients are the same in the balanced case ($n_1 = n_2$).

(5) The intercepts are different in the not balanced case ($n_1 \neq n_2$). But the coefficients associated to the predictors and the test for significance are valid. Thus, we can use a variable selection for regression in the classification problem.

(6) There is not an unique solution for the multi-class problem (**K > 2**). We have no longer the equivalence with the linear discriminant analysis.

References

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

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.