Classification I

Jing Qin

12/04/2022

Bull-data (again)

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
5	1300	48.7	1056	72.9	5	0.15	52.6	1525
1	1525	49.4	959	68.4	6	0.15	52.6	1565
1	1525	49.6	1083	75.8	6	0.30	54.6	1640
8	1850	53.1	964	70.8	8	0.10	55.5	1535
1	1500	49.5	963	69.4	6	0.35	53.1	1670
8	1825	53.0	1055	76.8	8	0.10	56.7	1526
5	1375	51.0	1002	72.1	7	0.25	51.9	1410
1	1400	47.6	974	69.7	5	0.15	51.9	1570
1	2250	51.9	1108	72.1	7	0.25	55.3	1575
8	2000	53.5	1175	74.5	8	0.10	57.4	1686
8	1725	51.4	1034	71.2	7	0.10	56.0	1655

Discrimination and classfication: same **rules** but *different* purposes

Discrimination (separation): to describe, graphically or algebraically, the differential features of objects from several known collections (populations). We try to find 'discriminants' whose numerical values are such that the collections are separated as much as possible.

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
5	1300	48.7	1056	72.9	5	0.15	52.6	1525
1	1525	49.4	959	68.4	6	0.15	52.6	1565
1	1525	49.6	1083	75.8	6	0.30	54.6	1640
8	1850	53.1	964	70.8	8	0.10	55.5	1535
1	1500	49.5	963	69.4	6	0.35	53.1	1670
8	1825	53.0	1055	76.8	8	0.10	56.7	1526
5	1375	51.0	1002	72.1	7	0.25	51.9	1410
1	1400	47.6	974	69.7	5	0.15	51.9	1570
1	2250	51.9	1108	72.1	7	0.25	55.3	1575
8	2000	53.5	1175	74.5	8	0.10	57.4	1686
8	1725	51.4	1034	71.2	7	0.10	56.0	1655

Discrimination and classfication:

same rules but *different* purposes

Classification (allocation): to sort objects into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes.

?Breed

SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
1300	48.7	1056	72.9	5	0.15	52.6	1525
1525	49.4	959	68.4	6	0.15	52.6	1565
1525	49.6	1083	75.8	6	0.30	54.6	1640
1850	53.1	964	70.8	8	0.10	55.5	1535
1500	49.5	963	69.4	6	0.35	53.1	1670
1825	53.0	1055	76.8	8	0.10	56.7	1526
1375	51.0	1002	72.1	7	0.25	51.9	1410
1400	47.6	974	69.7	5	0.15	51.9	1570
2250	51.9	1108	72.1	7	0.25	55.3	1575
2000	53.5	1175	74.5	8	0.10	57.4	1686
1725	51.4	1034	71.2	7	0.10	56.0	1655

We start with a bit easier setting: only two populations

Hemophilia A data set (Example 11.3)

Example: detection of hemophilia A carriers

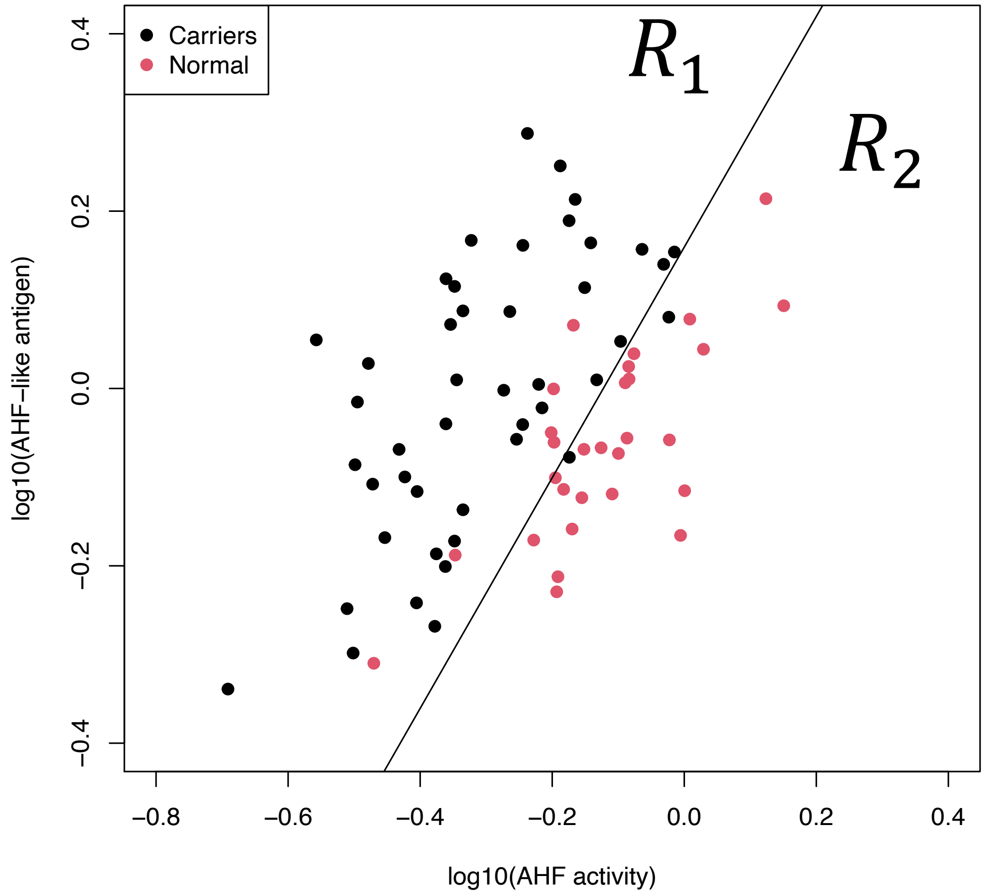
Classify people as normal, i.e. not carrying the hemophilia gene, or as obligatory carrier on the basis of the following blood sample measurements

$$X_1 = \log(\text{AHF activity}),$$

$$X_2 = \log(\text{AHF-like antigen}).$$

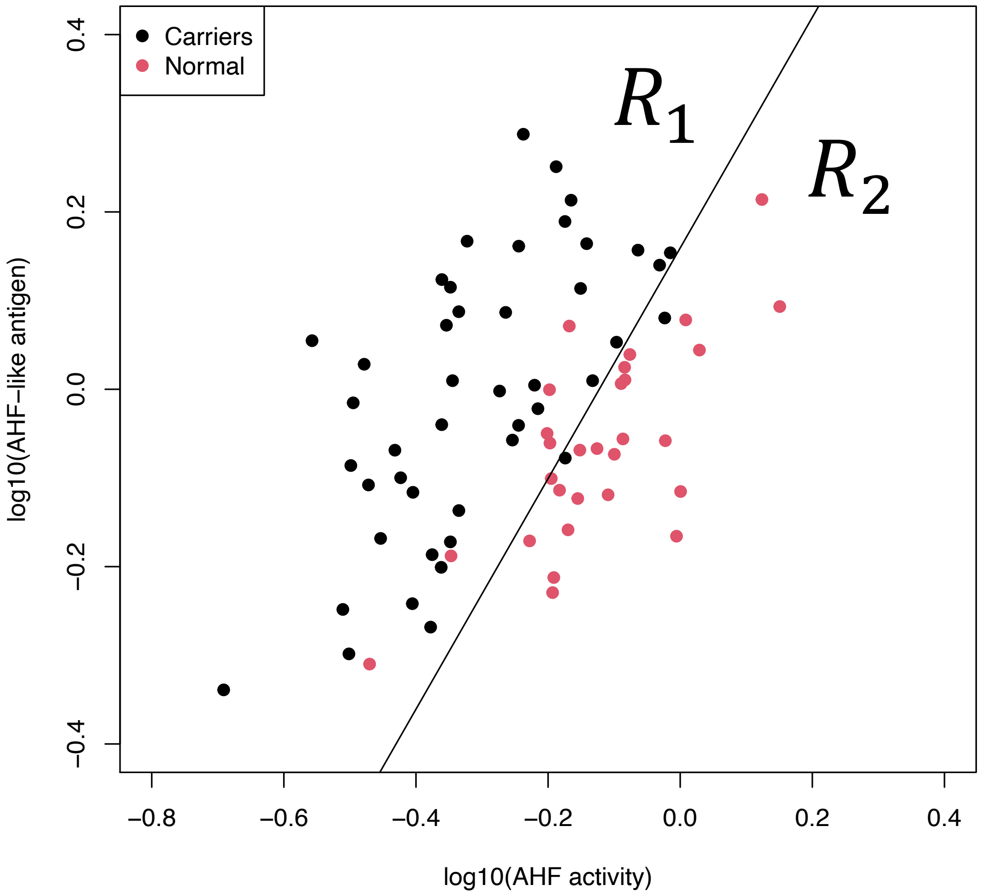
π_1 : truely in group 1 (carriers)
π_2 : truely in group 2 (normal)

R_1 : allocated in group 1 (carriers)
R_2 : allocated in group 2 (normal)



Some rules are designed to minimize the *risks*

Risk of
Misclassification #1



$$P(\,a\ normal\ observation\ is\ classified\ as\ carrier)$$

$$\|$$

$$P(\,observation\ is\ normal\ and\ classified\ as\ carrier)$$

$$\|$$

$$P(A\ \cap B) = P(B|A)\cdot P(A)$$

$$P(classified\ as\ carrier\ |observation\ is\ normal)$$

$$\times$$

$$P(\,observation\ is\ normal)$$

Some rules are designed to minimize the *risks*

Risk of Misclassification #2

$$P(\text{ a carrier observation is classified as normal})$$

$$\parallel$$

$$P(\text{ observation is carrier and classified as normal})$$

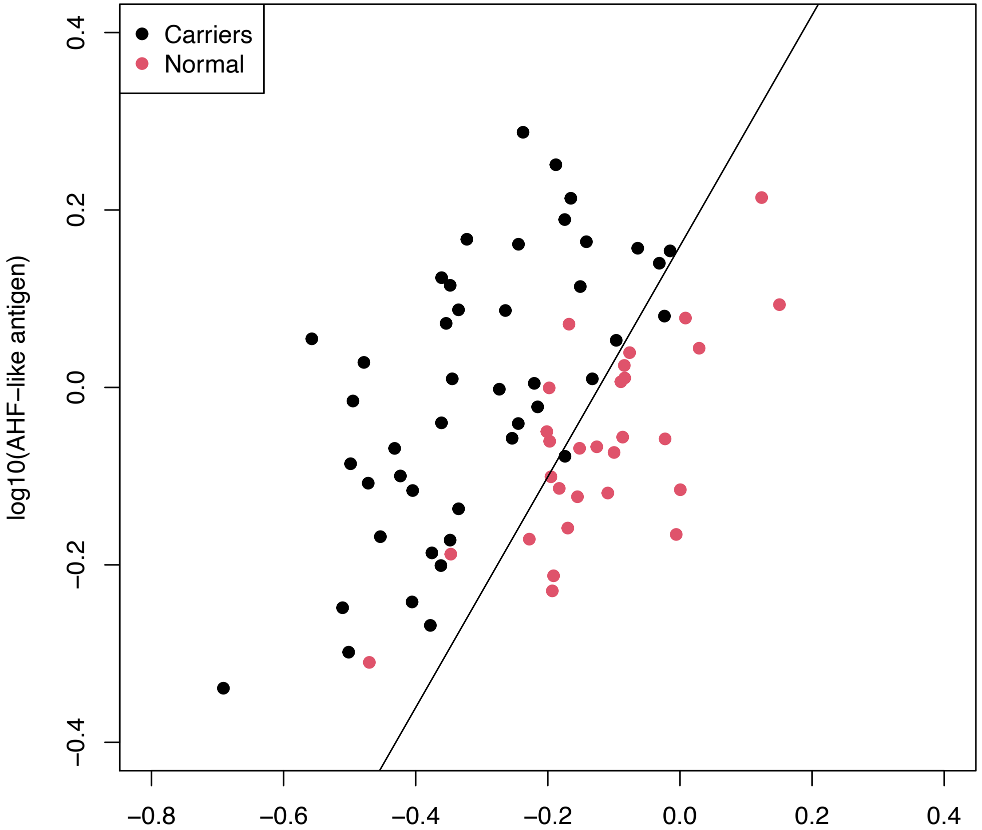
$$\parallel$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

$$P(\text{classified as normal} \mid \text{observation is carrier})$$

$$\times$$

$$P(\text{ observation is carrier})$$



Expected Cost of Misclassification (ECM)

$$\textit{Cost \#1} \times P(\textit{classified as carrier} \mid \textit{observation is normal}) \times P(\textit{observation is normal})$$

$$+$$

$$\textit{Cost \#2} \times P(\textit{classified as normal} \mid \textit{observation is carrier}) \times P(\textit{observation is carrier})$$

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \tag{11-5}$$

$$P(\textit{classified as Group 2} \mid \textit{observation is supposed to be in Group 1})$$

Expected Cost of Misclassification (ECM)

$$Cost\ \#1 \times P(classified\ as\ carrier \mid observation\ is\ normal)$$
$$\times$$
$$P(\ observation\ is\ normal)$$

$$+$$

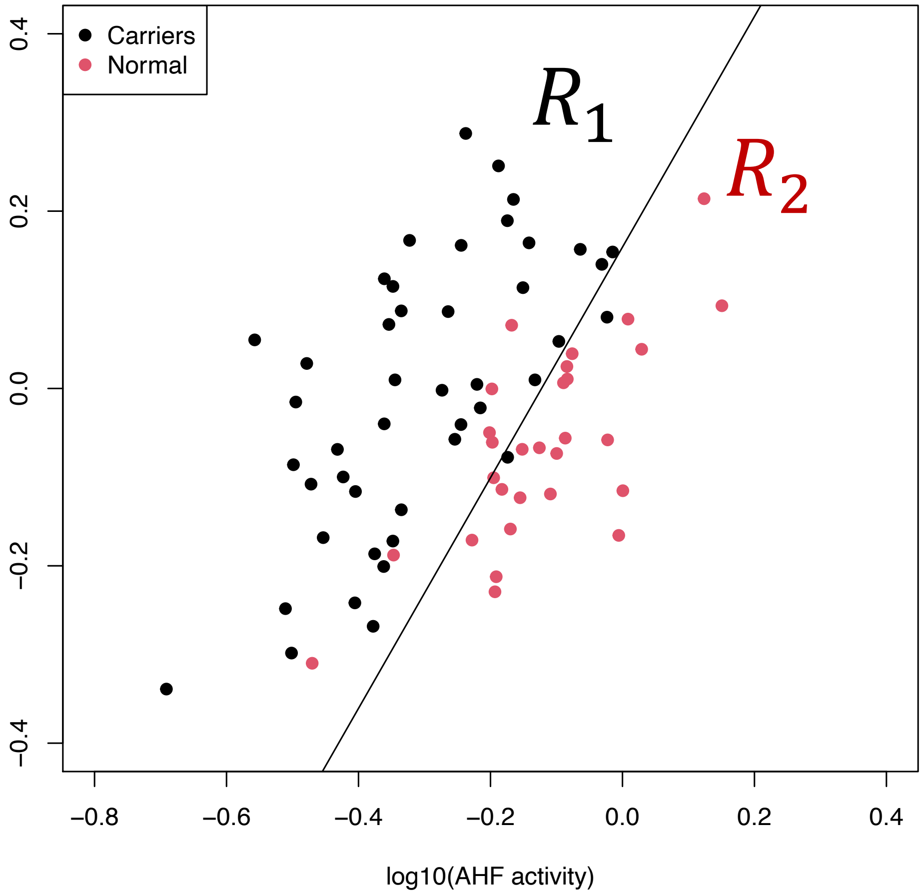
$$Cost\ \#2 \times P(classified\ as\ normal \mid observation\ is\ carrier)$$
$$\times$$
$$P(\ observation\ is\ carrier)$$

$$\text{ECM} = c(2 \mid 1)P(2 \mid 1)p_1 + c(1 \mid 2)P(1 \mid 2)p_2 \tag{11-5}$$

$$P(classified\ as\ Group\ 2 \mid observation\ is\ supposed\ to\ be\ in\ Group\ 1)$$

$$\text{ECM} = c(2|1)\underline{P(2|1)}p_1 + c(1|2)\underline{P(1|2)}p_2 \tag{11-5}$$

Result 11.1. The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:



$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\\text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\\text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\\text{probability}\\\text{ratio}\end{pmatrix} \tag{11-6}$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\\text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\\text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\\text{probability}\\\text{ratio}\end{pmatrix}$$

Special Cases of Minimum Expected Cost Regions

(a) $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

TPM

Total probability of misclassification rule

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \tag{11-7}$$

(c) $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$ (equal prior probabilities and equal misclassification costs)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Example 11.2

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

Example 11.2

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

$\boldsymbol{x_0}$ should be allocated to Group 1, R_1

If, normally distributed (Test it yourself!)

$$N(\boldsymbol{\mu}_1, \Sigma_1) \rightarrow f_1(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_1|} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)/2\}$$

The (R_1, R_2) minimize ECM is

$$R_1 = \left\{ \boldsymbol{x} \,\middle|\, \frac{\boxed{f_1(\boldsymbol{x})}}{\boxed{f_2(\boldsymbol{x})}} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right\}$$

$$N(\boldsymbol{\mu}_2, \Sigma_2) \rightarrow f_2(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_2|} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)/2\}$$

If, normally distributed, further with MASS in R

$\Sigma_1 = \Sigma_2$?

Homogeneous?

$\Sigma_1 = \Sigma_2 = \Sigma$

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{11-18}$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \tag{11-17}$$

linear discriminant analysis (R cmd `lda()` and `predict()`)

$\Sigma_1 \neq \Sigma_2$

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \tag{11-29}$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

quadratic discriminant analysis (R cmd `qda()` and `predict()`)

If **not** normally distributed, use logistic regression model §11.7

- The method depends on a logistic regression model

$$\ln\left(\frac{p}{1-p}\right) \qquad value = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  in which coefficients

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$$

  are determined based on maximum likelihood theory.

- In practice, the model including $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ can be constructed with R cmd

```
glm('group' ~ 'predictors', data, family=binomial(link="logit"))
```

- Classification criterion: Allocate $\boldsymbol{x}$ to group 1 if the estimated posterior probability

$$\hat{P}(1|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{value}}{1 + e^{value}} \geq 1/2$$

Well, we have LDA, QDA and logistic…so compare

Comparison?

APER. apparent error rate

$$\frac{\#\text{ of data misclassified}}{\text{size of data}}$$

more criterion coming later

Confusion matrix   table()

		predicted π_1	predicted π_2	
Actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

LDA

27	3
8	37

QDA

27	3
8	37

Logistic

25	5
4	41

$$\frac{9}{75} = 12\%$$

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$?
Homogeneous in general: Box's M-test §6.6

Assume g different groups with distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and there is *independence* between the observations belonging to different groups. We are interested in testing

$$\begin{aligned} H_0 &: \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g, \\ H_1 &: \quad \text{at least two } \boldsymbol{\Sigma}_i \text{ not equal }, \end{aligned}$$

at the significance level α.

Box's M-test (maximum likelihood test)

- Test statistic: $M = -2\ln\Lambda$

$$\boxed{M = (n-g)\ln|\boldsymbol{S}_{\text{pooled}}| - \sum_{\ell=1}^{g}(n_\ell - 1)\ln|\boldsymbol{S}_\ell|,}$$

In which, likelihood ratio $\Lambda = \prod_{\ell=1}^{g}\left(\frac{|\boldsymbol{S}_\ell|}{|\boldsymbol{S}_{\text{pooled}}|}\right)^{(n_\ell-1)/2}$,

with

$$\boldsymbol{S}_\ell = \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)',$$

$$\boldsymbol{S}_{\text{pooled}} = \frac{1}{n-g}\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)',$$

where $n = \sum_{\ell=1}^{g} n_\ell$.

Box's M-test

Then, under H_0

$$(1-u)M \dot{\sim} \chi^2_\nu$$

$$u = \left[\sum_{\ell=1}^{g} \frac{1}{n_\ell - 1} - \frac{1}{n-g}\right] \frac{2p^2+3p-1}{6(p+1)(g-1)}. \qquad \nu = \frac{1}{2}\, p(p+1)(g-1).$$

Reject H_0, if $(1-u)M > \chi^2_v(\alpha)$

Try it out with the Example 11.3, i.e. hemophilia data set