多變量分析 HW4

劉昱維, 吳冠瑋, 廖永賦, 謝靖惟, 黃奎鈞

Contents

11.1

- a) $\hat{y} = (\bar{x_1} - \bar{x_2})' S_{pooled}^{-1} x = \hat{a}' x \qquad S_{pooled}^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \qquad \begin{bmatrix} -2 & -2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} x = -2x_1$
- b) $\hat{m} = \frac{1}{2}(\hat{y_1} + \hat{y_2}) = -8 \qquad -2(2) = -4 > -8 \Rightarrow assign\, to\, \pi_1$

11.4

- a) $(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1}) = 0.5 \qquad \Rightarrow assign\, to\, \pi_1\, if\, \frac{f_1(x)}{f_2(x)} \geq 0.5,\ otherwise\, assign\, to\, \pi_2$
- b) $\frac{f_1(x)}{f_2(x)} = \frac{0.3}{0.5} = 0.6 > 0.5 \Rightarrow assign\, to\, \pi_1$

11.19

- b) $\hat{y_0} = \hat{a}' x_0 = -0.33x_1 + 0.67x_2 \qquad where\ \hat{m} = 4.5 \qquad assign\, to\, \pi_1\, if\ -0.33x_1 + 0.67x_2 - 4.5 \geq 0,\ otherwise\, assign\, to\, \pi_2$

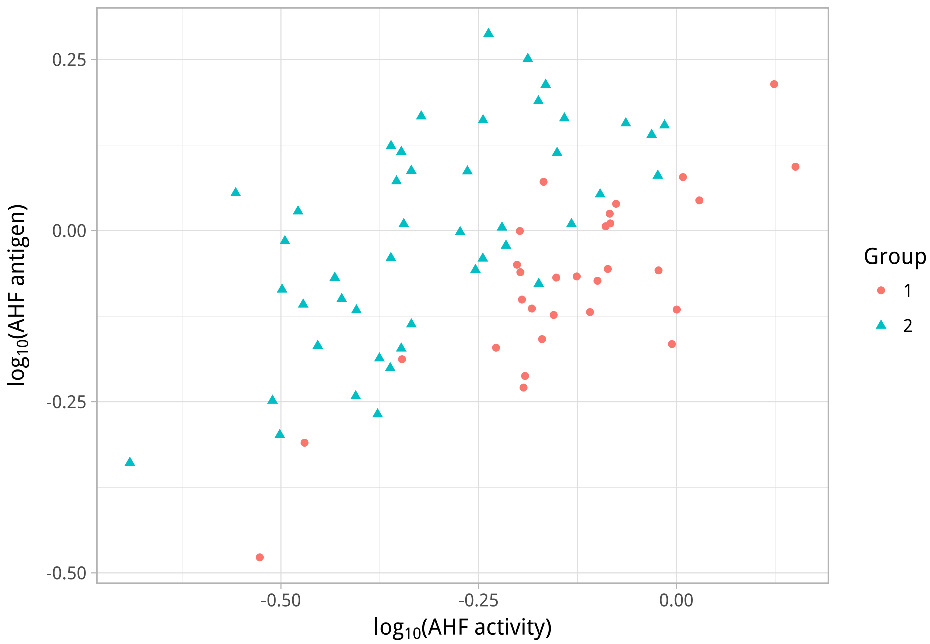
	π_1			π_2	
obs.	$0.33x_1 + 0.67x_2 4.5$	class.	obs.	$0.33x_1 + 0.67x_2 4.5$	class.
1	2.83	π_1	1	-1.5	π_2
2	0.83	π_1	2	0.5	π_1
3	-0.17	π_2	3	-2.5	π_2

- c) $D_i^2(x) = (x - \bar{x}_i)' S_{pooled}^{-1} (x - \bar{x}_i)$

		π_1			π_2
obs.	D_1^2	D_2^2	class.	obs.	D_1^2
1	1.33	7	π_1	1	4.33
2	1.33	3	π_1	2	0.33
3	1.33	1	π_2	3	6.33

11.32

(a) Bivariate plot



The bivariate plot doesn’t seem to fit well into a bivariate normal distribution.

SAS `pool=test`

集區類別內共變異數矩陣、 DF = 73		
變數	log10_AHF_Act	log10_AHF_Anti
log10_AHF_Act	0.0226370923	0.0154313022
log10_AHF_Anti	0.0154313022	0.0216058018

discrimination analysis

DISCRIM 程序

共變異數內矩陣的均齊性檢定

卡方	自由度	Pr > ChiSq
5.338277	3	0.1486

因為卡方值在 0.1 層級不顯著，所以會在判別函數中使用集區共變異數矩陣。

參考: Morrison , D.F. (1976) 多變量統計法第 252 頁。

Test not significant. The covariance of the two groups are assumed to be equal.

SAS `manova`

DISCRIM 程序

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=35					
統計值	值	F 值	分子自由度	分母自由度	Pr > F
Wilks' Lambda	0.46994422	40.60	2	72	<.0001
Pillai's Trace	0.53005578	40.60	2	72	<.0001
Hotelling-Lawley Trace	1.12791213	40.60	2	72	<.0001
Roy's Greatest Root	1.12791213	40.60	2	72	<.0001

線性判別函數: Group		
變數	1	2
常數	-0.41072	-3.97019
log10_AHF_Act	-6.82377	-26.14277
log10_AHF_Anti	1.27017	18.39440

Means of two groups are not equal.

(b) linear discriminant function

$$\begin{aligned}\hat{y} &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S_{pool}^{-1}\, \boldsymbol{x}_0 \\ &= \boldsymbol{a}^T \boldsymbol{x}_0 \\ &= \begin{pmatrix} 19.319 & -17.124 \end{pmatrix} \boldsymbol{x}_0 \end{aligned} \tag{1}$$

Then allocate $\boldsymbol{x}_0$ to π_1 if:

$$\begin{aligned}\hat{y} &\geq \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S_{pool}^{-1}\, (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \\ &= \frac{1}{2}\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = -3.559 \end{aligned} \tag{2}$$

where, $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} -0.135 \\ -0.078 \end{pmatrix}, \bar{\boldsymbol{x}}_2 = \begin{pmatrix} -0.308 \\ -0.006 \end{pmatrix}$

The linear discriminant function is:

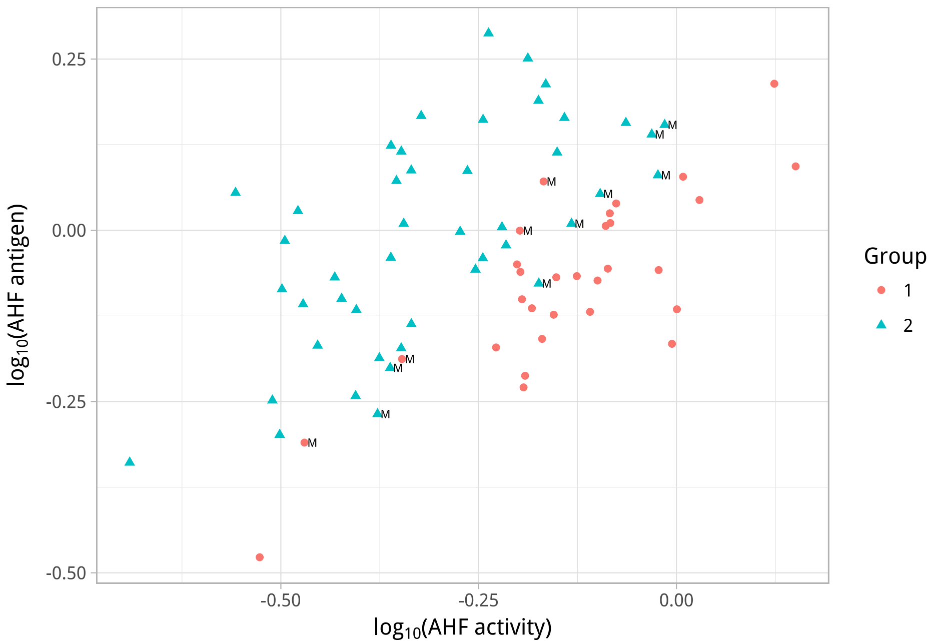
$$19.319 x_{0,1} - 17.124 x_{0,1} + 3.559 \tag{3}$$

The confusion matrix constructed with the holdout procedure is

	Group 1	Group 2
1	26	4
2	8	37

, and the estimated error rate is **0.16**.

The misclassified observations are **No. 3, 5, 7, 17, 32, 35, 58, 62, 63, 64, 67, 69**, labeled "M" in the plot below.



(c)

By eq. (1) and eq. (2), $\hat{y}_0 = 2.614 \geq -3.559$. Hence, it is allocated to **Group 1**.

(d)

Classification rule based on posterior probabilities is equivalent to classification rule based on minimizing TPM.

Since the prior probabilities are assumed to be equal, the posterior probabilities are calculated as:

$$
\begin{aligned}
p(\pi_1|\boldsymbol{x}_0) &= \frac{f_1(\boldsymbol{x}_0)}{f_1(\boldsymbol{x}_0) + f_2(\boldsymbol{x}_0)} = 0.9608785 \\
p(\pi_2|\boldsymbol{x}_0) &= 1 - p(\pi_1|\boldsymbol{x}_0) = 0.0389789
\end{aligned}
\tag{4}
$$

, where the densities $f_1(\boldsymbol{x}_0)$ and $f_2(\boldsymbol{x}_0)$ are assumed to be normal and are estimated using $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \boldsymbol{S}_1, \boldsymbol{S}_2$.

By $p(\pi_1|\boldsymbol{x}_0) > p(\pi_2|\boldsymbol{x}_0)$, $\boldsymbol{x}_0$ is classified as **Group 1**.

(e)

By eq. (1), eq. (2), and c, the linear discriminant score is calculated as $\hat{y}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S_{pool}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = 6.173$.

(f)

Assume $p_1 = 0.75$ and $p_2 = 0.25$, then allocate $\boldsymbol{x}_0$ to π_1 if:

$$\begin{aligned}\hat{y} &\geq \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T S_{pool}^{-1}\,(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + ln(\frac{c(1|2)p_2}{c(2|1)p_1}) \\ &= \frac{1}{2}\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + ln(\frac{0.25}{0.75}) = -4.658\end{aligned} \tag{5}$$

where, $\bar{\boldsymbol{x}}_1 = \begin{pmatrix}-0.135 \\ -0.078\end{pmatrix}$, $\bar{\boldsymbol{x}}_2 = \begin{pmatrix}-0.308 \\ -0.006\end{pmatrix}$

The linear discriminant function is:

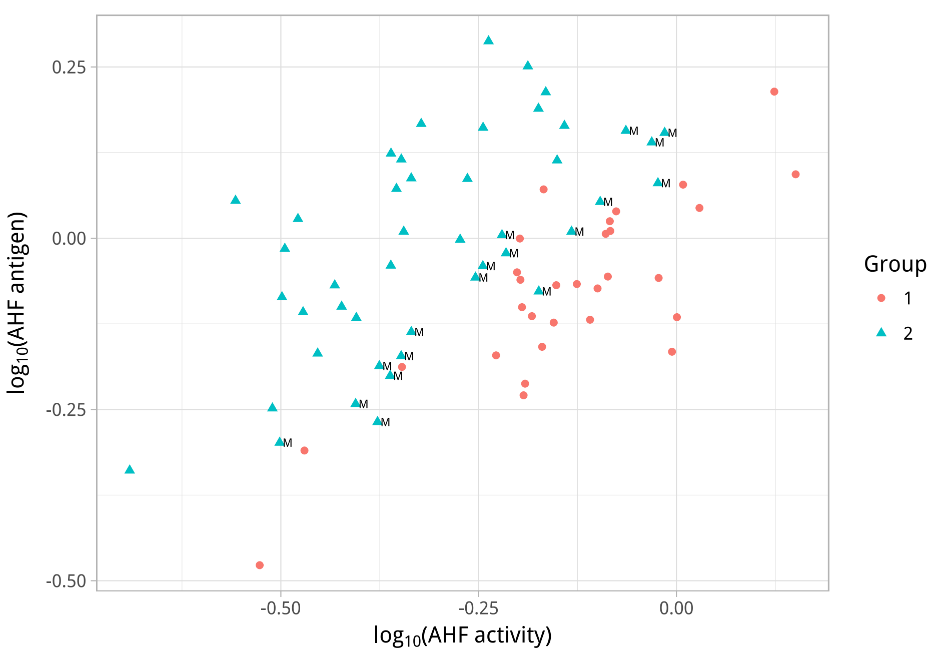
$$19.319x_{0,1} - 17.124x_{0,1} + 4.658 \tag{6}$$

The confusion matrix constructed with the holdout procedure is

	Group 1	Group 2
1	30	0
2	18	27

, and the estimated error rate is **0.24**.

The misclassified observations are **No. 32, 34, 35, 39, 47, 51, 54, 55, 57, 58, 60, 61, 62, 63, 64, 67, 69, 73**, labeled "M" in the plot below.



(g)

By eq. (1) and eq. (5), $\hat{y}_0 - [\frac{1}{2}\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + ln(\frac{0.25}{0.75})] = 7.272 > 0$. Hence, it is allocated to **Group 1**.

(h)

Classification rule based on posterior probabilities is equivalent to classification rule based on minimizing TPM.

The posterior probabilities are calculated as:

$$\begin{aligned} p(\pi_1|\boldsymbol{x}_0) &= \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)} = 0.9876759 \\ p(\pi_2|\boldsymbol{x}_0) &= 1 - p(\pi_1|\boldsymbol{x}_0) = 0.0133553 \end{aligned} \tag{7}$$

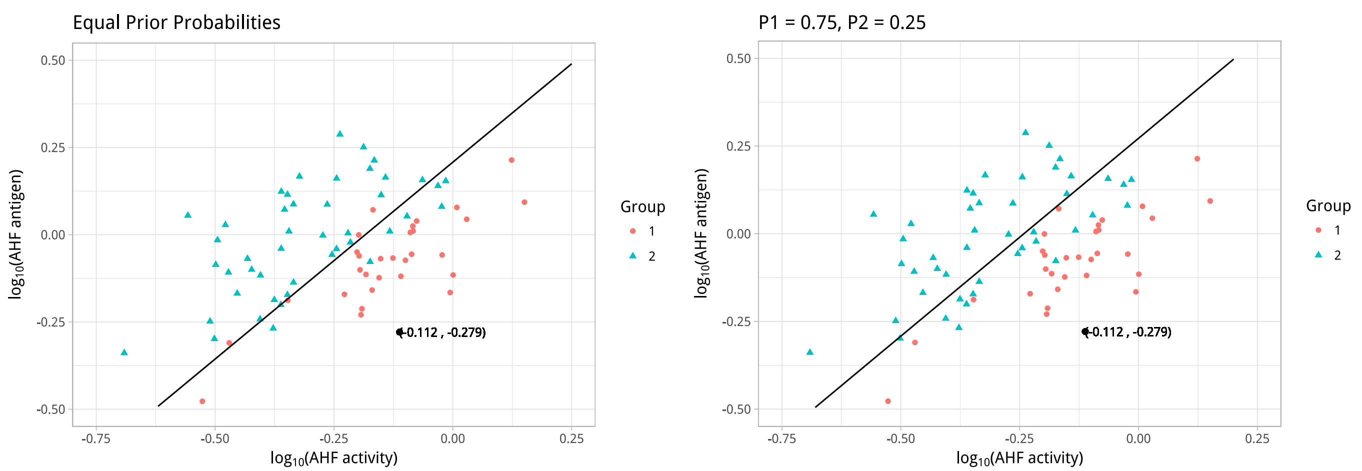
, where the densities $f_1(\boldsymbol{x}_0)$ and $f_2(\boldsymbol{x}_0)$ are assumed to be normal and are estimated using $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \boldsymbol{S}_1, \boldsymbol{S}_2$.

By $p(\pi_1|\boldsymbol{x}_0) > p(\pi_2|\boldsymbol{x}_0)$, $\boldsymbol{x}_0$ is classified as **Group 1**.

(i)

By eq. (1), eq. (5), and g, the linear discriminant score is calculated as $\hat{y}_0 - [\frac{1}{2}\boldsymbol{a}^T(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + ln(\frac{0.25}{0.75})] = 7.271934$.

(j)



When the prior probability p_1 changes from 0.5 to 0.75, the discriminant function shifts parallelly to the upper-left direction.