11. 판별분석(Discriminant Analysis) 및 분류법(Classification)

11.1 개념 및 목적

두 개 이상의 모집단에서 추출된 표본들이 지니고 있는 정보를 이용하여 이 표본들이 어느 모집단에서 추출된 것인지를 결정해 줄 수 있는 기준을 찾는 분석법

Discriminant(판별법)

미리 알려진 여러 개의 그룹(모집단)에서 관측된 개체들을 각 그룹별로 구분하되, 가능한 한 각 개체들을 원래 속해 있던 모집단으로 판별해 주는 방법으로, 각 개체와 그룹의 중심(모평균)과의 거리를 계산하여 가장 가까운 그룹으로 판별해 준다.

Classification(분류법)

새로 관측된 개체를 사전에 알고 있던 여러 개의 그룹 중에서 하나의 그룹으로 분류하는 방법으로, 잘못 분류되는 경우가 최소가 되도록 한다.

예 11.1 장래에 대한 불확실성으로 인해 결정하기 어려운 경우

- 건실한 기업과 부실기업을 구분하는 문제 (Sound firm or bankrupcy)
- 의사 지망생 중 의사로서의 가능성이 있는지여부를 사전에 판단해야 하는 경우 (likely to become M.D. or unlikely to become M.D.)

예 11.2 완전한 정보를 얻기 위해서는 대상이 되는 개체를 파괴해야 하는 경우

- 가전제품의 수명, 전구의 수명 또는 건전지의 수명들을 알아보기 위해서는 이들의 수명이 다할 때까지 사용해야 하므로, 사전에 비교적 관측하기 쉬운 방법을 이용하여 불량여부를 판단

예 11.3 정보를 얻기 어렵거나 비용이 많이 드는 경우

- Federalist papers의 저자를 판단하는 문제(James Madison or Alexander Hamilton)
- Medical problems(requires an expensive operations)

판별함수(discriminant function)

판별의 기준으로는 잘못 분류될 확률 또는 잘못 분류됐을 경우 발생할 수 있는 비용 등을 최소로 하는 확률변수들의 함수

분류함수(classification function)

어느 모집단에서 추출된 것인지를 모르는 새로운 표본이 관측되었을 때 이 표본을 여러 모집단 중에서 어느 하나의 모집단으로 분류(classification)해 주기 위해 분류의 기준을 사용되는 함수. 판별분석에서 구해진 판별함수를 분류함수로 사용하는 것이 보통

예제 11.1 (두 모집단의 경우)

운전용 잔디 깍는 기계(riding-mower)를 생산하는 회사에서는 판매를 촉진하기 위해 x_1 = 소득(income) 과 x_2 = 대지크기(lot size)에 따라 기계를 살 가능성이 있는 가구인지의 여부를 판별하고자 한다.

잔디 깍는 기계를 소유한 가구 $n_1 = 12$

잔디 깍는 기계를 소유하지 않은 가구 $n_2 = 12$

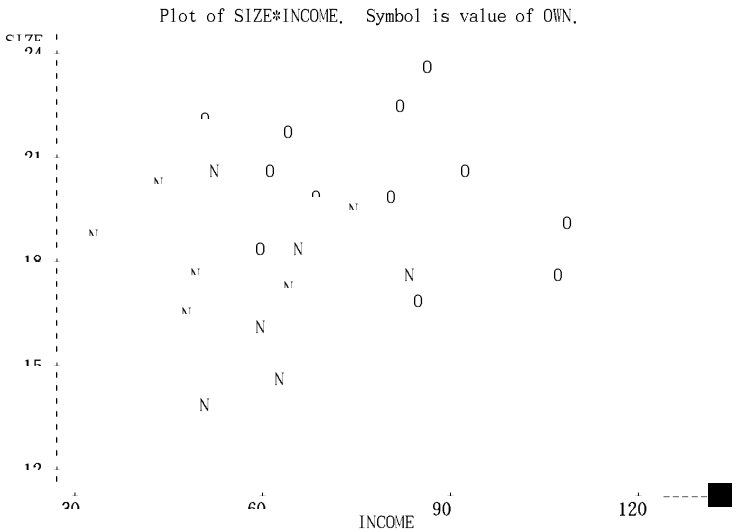
랜덤 추출하여 조사한 결과표

<표 11.1> 수입과 대지크기의 자료

소유자(owner)		비소유자(nonowner)	
x_1	x_2	x_1	x_2
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

단위 : $x_1 - 1000$불, $x_2 - 1000ft^2$

$x_1 = 60.5,\ x_2 = 19.5$인 가구와 $x_1 = 79.0,\ x_2 = 18.0$인 가구는 구매 가능성에 대하여 어떻게 분류되겠는가?



<그림 11.1> 소득(income)과 대지크기(size)의 산점도

예제 11.2 (세 모집단의 경우)

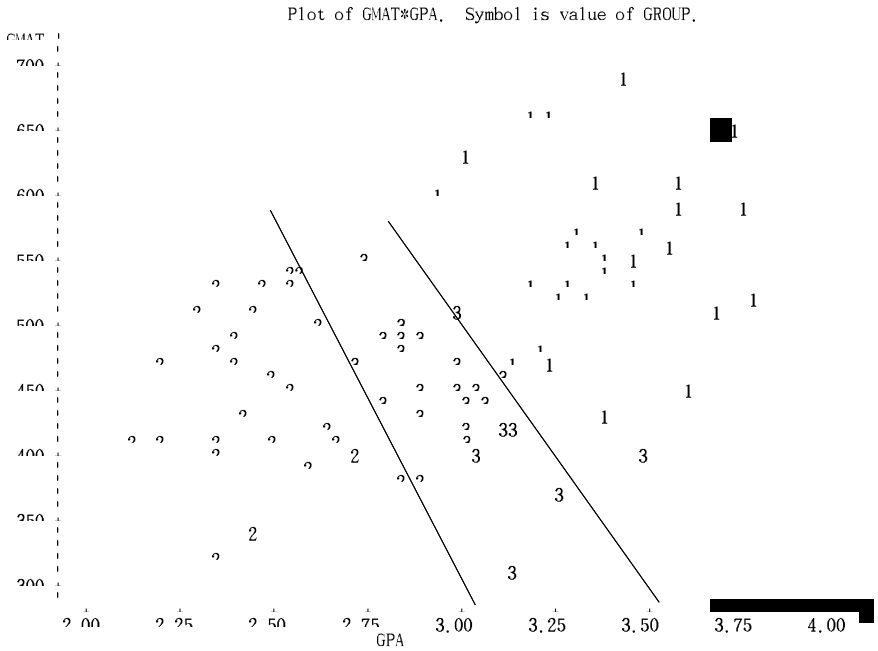
경영대학원의 지원자들의 입학허가여부를 판정의 기준변수

x_1 = GPA(Undergraduate Grade Point Average)

x_2 = GMAT(Graduate Management Aptitude Test)

π_1 : 입학허가, π_2 : 입학불허, π_3 : 보류

최근의 지원자들의 관측값의 산점도



<그림 11.2> 경영대학원 지원자 자료의 산점도

새로운 지원자의 성적이 GPA= 3.21, GMAT= 479일 때 입학을 허가해야 하겠는가?

좋은 분류법(optimal classification rule)을 선택하는 기준

1) 잘못 분류될 확률을 최소로 하는 방법

2) 모집단의 크기가 다른 경우 이를 사전 확률로 이용

3) 잘못 분류될 경우의 비용이 현저히 다른 경우 이를 반영

용어 :

$f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$: 모집단 π_1과 π_2의 밀도함수

R_1, R_2 : π_1과 π_2로 분류되는 개체 $\boldsymbol{x}$의 집합 ($\Omega = R_1 + R_2$)

p_1, p_2 : π_1과 π_2의 사전확률 ($p_1 + p_2 = 1$)

모집단 π_1에 속한 개체를 π_2로 잘못 분류할 확률

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}$$

모집단 π_2에 속한 개체를 π_1으로 잘못 분류할 확률

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

$P(\pi_1$으로 옳게 분류될 확률$) = P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1$

$P(\pi_1$으로 잘못 분류될 확률$) = P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2$

$P(\pi_2$로 옳게 분류될 확률$) = P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2$

$P(\pi_2$로 잘못 분류될 확률$) = P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1$

$c(2|1)$: 모집단 π_1에 속한 개체를 π_2로 잘못 분류할 경우의 비용

$c(1|2)$: 모집단 π_2에 속한 개체를 π_1으로 잘못 분류할 경우의 비용

===> 잘못 분류할 경우의 기대비용

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

정리 11.1 (minimum *ECM* rule)

ECM 을 최소로 하는 R_1과 R_2는 다음과 같다.

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

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

(minimum *TPM* rule)

잘못 분류될 경우의 비용을 무시할 수 있다면

Total Probability of Misclassification(*TPM*)

$$TPM = P(2|1)p_1 + P(1|2)p_2$$

$$= p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

TPM 을 최소로 하는 R_1과 R_2

$$R_1 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqq \frac{p_2}{p_1}, \quad R_2 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \leqq \frac{p_2}{p_1}$$

(maximum posterior probability)

새로 관측된 개체 $\boldsymbol{x_0}$을 사후확률이 큰 모집단으로 분류하는 방법

$$P(\pi_1|\boldsymbol{x_0}) = \frac{p_1 f_1(\boldsymbol{x_0})}{p_1 f_1(\boldsymbol{x_0}) + p_2 f_2(\boldsymbol{x_0})}, \quad P(\pi_2|\boldsymbol{x_0}) = \frac{p_2 f_2(\boldsymbol{x_0})}{p_1 f_1(\boldsymbol{x_0}) + p_2 f_2(\boldsymbol{x_0})}$$

즉, $\frac{P(\pi_1|\boldsymbol{x_0})}{P(\pi_2|\boldsymbol{x_0})} = \frac{p_1 f_1(\boldsymbol{x_0})}{p_2 f_2(\boldsymbol{x_0})} \geqq 1$ **이면** π_1**으로 분류**

$$\Leftrightarrow \frac{f_1(\boldsymbol{x_0})}{f_2(\boldsymbol{x_0})} \geqq \frac{p_2}{p_1}$$: **minimum** *TPM* **rule과 동일**

11.2 이론적 배경

(1) 두 모집단의 경우

1) Fisher의 선형판별함수(사전확률이 동일)

두 모집단 : π_1, π_2

각 그룹에 대한 분포 가정은 필요 없으나 산포는 같다고 가정

즉, 정규가정이 필요 없으나 두 모집단의 공분산 행렬은 동일하다고 가정

$$\boldsymbol{\Sigma} = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})'$$

선형판별함수 : 모집단들 사이의 거리가 가장 최대가 되는 선형결합식을 이용하여 다변량 관측값 $\boldsymbol{x}$를 일변량 관측값 y로 변환

y : 관측값 $\boldsymbol{x}' = (x_1, x_2, \cdots, x_p)$의 일차결합

$$y = \ell_1 x_1 + \ell_2 x_2 + \cdots + \ell_p x_p = \boldsymbol{\ell}'\boldsymbol{x}, \quad \text{단, } \boldsymbol{\ell} = (\ell_1, \ell_2, \cdots, \ell_p)$$

① 계수벡터 $\boldsymbol{\ell}$의 결정

μ_{1y} : π_1에 속하는 y의 평균

μ_{2y} : π_2에 속하는 y의 평균

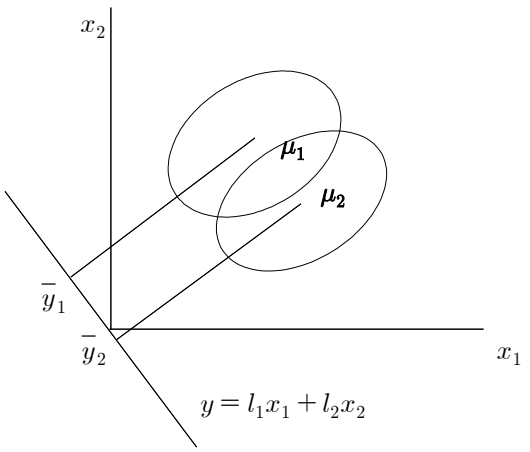
μ_{1y}와 μ_{2y} 사이의 거리가 최대로 되도록 $\boldsymbol{\ell}$을 구한다.

즉, 다음의 비가 최대가 되도록 한다.

$$\frac{y\text{의 표본평균 사이의 거리의 제곱}}{y\text{의 표본분산}}$$

$$= \frac{(\bar{y}_1 - \bar{y}_2)^2}{S_y^2} = \frac{(\hat{\boldsymbol{\ell}}'\bar{\boldsymbol{x}}_1 - \hat{\boldsymbol{\ell}}'\bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{\ell}}'\boldsymbol{S}_p\hat{\boldsymbol{\ell}}} = \frac{(\hat{\boldsymbol{\ell}}\boldsymbol{d})^2}{\hat{\boldsymbol{\ell}}'\boldsymbol{S}_p\hat{\boldsymbol{\ell}}}$$

단, $\boldsymbol{d} = \bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2$. 즉, 표본분산(within sample variation)과 표본 간의 분산(between sample variation)의 비가 최대가 되도록 $\boldsymbol{\ell}$을 구한다.



두 모집단이 동일한 공분산행렬 $\boldsymbol{\Sigma}$ 를 갖는다는 가정하에서 $\boldsymbol{\ell}$ 을 구하면

$$\boldsymbol{\ell} = \Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$

단, $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$: π_1, π_2 의 모평균벡터가 되어

$$y = \boldsymbol{\ell}'\boldsymbol{x} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\Sigma^{-1}\boldsymbol{x}$$

: Fisher의 선형판별함수(Fisher's linear discriminant function)

② μ_i 와 Σ : 미지인 경우

$\boldsymbol{\mu}$와 Σ 에 대한 추정량인 표본평균과 표본 공분산 행렬을 대입하여 사용

$\boldsymbol{x}_{11}, \boldsymbol{x}_{12}, \cdots, \boldsymbol{x}_{1n_1}$: 첫 번째 모집단에 속한 개체들의 관측값

$\boldsymbol{x}_{21}, \boldsymbol{x}_{22}, \cdots, \boldsymbol{x}_{2n_2}$: 두 번째 모집단에 속한 개체들의 관측값

Fisher의 표본판별함수

$$y = \hat{\boldsymbol{\ell}}'\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'S_p^{-1}\boldsymbol{x}$$

$$\bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}, \quad i = 1,2 \ : \ \text{표본평균}$$

S_p : $\boldsymbol{\Sigma}$ 의 합동추정량

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1 + n_2 - 2}$$

단, $S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)'$

$$S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)'$$

③ Fisher의 판별함수를 이용한 분류방법

기준점 : μ_{1y} 와 μ_{2y} 의 중앙값

$$m = \frac{1}{2}(\mu_{1y} + \mu_{2y}) = \frac{1}{2}(\boldsymbol{\ell}'\boldsymbol{\mu}_1 + \boldsymbol{\ell}'\boldsymbol{\mu}_2)$$

$$= \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$$

표본에서의 분류 기준점

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)\, S_p^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)$$

즉, 새로운 개체 $\boldsymbol{x}_0$가 주어질 때

$$y_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\, S_p^{-1}\boldsymbol{x}_0$$

를 계산하고

$y_0 \geq \hat{m}$ 이면 π_1으로 분류

$y_0 \leq \hat{m}$ 이면 π_2로 분류

예제 11.1(계속) 평균과 공분산행렬을 구하면

$$\bar{\boldsymbol{x}}_1 = \begin{bmatrix} 79.48 \\ 20.27 \end{bmatrix} \quad \bar{\boldsymbol{x}}_2 = \begin{bmatrix} 57.40 \\ 17.63 \end{bmatrix} \quad S_p = \begin{bmatrix} 276.67 & -7.20 \\ & 4.27 \end{bmatrix}$$

Fisher의 선형판별함수

$$y = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} \boldsymbol{x} = 0.10x_1 + 0.79x_2$$

분류기준점 $\hat{m}$

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = (0.10 \quad 0.79)\begin{pmatrix} 68.44 \\ 18.95 \end{pmatrix}$$

$$= 21.8$$

분류를 원하는 두 개체에 대한 판별함수의 값

$$y_1 = 0.10(60.5) + 0.79(19.5) = 21.46 < 21.8$$

$$y_2 = 0.10(79.0) + 0.79(18.0) = 22.12 > 21.8$$

$\Rightarrow$ (60.5, 19.5)는 비소유그룹으로 분류
(79.0, 18.0)은 소유그룹으로 분류

④ 사전확률의 정보를 이용하는 판별함수

p_1, p_2 : 두 모집단 π_1, π_2의 사전확률(prior probability)

분류 기준값

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_p^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) + \ln \frac{p_2}{p_1}$$

즉, p_1이 p_2 보다 클수록 $\hat{m}$ 의 값은 작아지므로 개체는 π_1으로 분류될 가능성이 커지게 된다.

2) 두 다변량 정규모집단의 경우

$f_1(X) \sim N_p(\mu_1, \Sigma_1)$: 사전확률 p_1
$f_2(X) \sim N_p(\mu_2, \Sigma_2)$: 사전확률 p_2

① $\Sigma_1 = \Sigma_2 (= \Sigma)$ 인 경우

정리 11.2 minimum *ECM* rule (Expected Cost of Misclassification)

잘못 분류될 기대 비용을 최소화하는 방법 이용

$$(\mu_1 - \mu_2)' \Sigma^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

이면 개체 x_0를 π_1으로 분류

$$W = (\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

$$= (\mu_1 - \mu_2)'\Sigma^{-1}[x_0 - \frac{1}{2}(\mu_1 + \mu_2)]$$

: **Anderson의 분류함수**

$\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$이면 두 다변량 정규 모집단의 경우 **Fisher의 방법**과 **minimum** *ECM* **rule**은 동일하다.

$$W = (\mu_1 - \mu_2)'\Sigma^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

$$= \hat{l}'x_0 - \hat{m}$$

② 두 모집단의 공분산행렬이 다를 경우($\Sigma_1 \neq \Sigma_2$ 인 경우)

정리 11.3 이차분류함수(quadratic classification rule)

$$-\frac{1}{2}x_0'(\Sigma_1^{-1} - \Sigma_2^{-1})x_0 - (\mu_1'\Sigma_1^{-1} - \mu_2'\Sigma_2^{-1})x_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

이면 x_0을 π_1 으로 분류, 그렇지 않으면 π_2로 분류

단, $k = \frac{1}{2}ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) - \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2)$

유의할 점 : 정규가정에 매우 민감

정규가정을 만족하지 않을 경우의 해결방법

i) 정규변환을 시행하고, 공분산행렬의 동일성검정을 실시

ii) 분포형태에 관계없이 선형함수(또는 이차함수)를 이용

가능하면 표본을 training 표본과 validation 표본으로 나누어 분류함수를 비교해 볼 것

(2) $g(\geq 2)$ 모집단인 경우

1) 분포가정이 없는 경우

잘못 분류될 기대 비용을 최소화하는 방법 이용

$$P(k|i) = P(\pi_k|\pi_i) = \int_{R_k} f_i(x)dx, \quad k = 1, 2, \ldots, g$$

$$P(i|i) = 1 - \sum_{k \neq i}^{g} P(k|i)$$

$$ECM(1) = P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1)$$

$$= \sum_{k=2}^{g} P(k|1)c(k|1)$$

: π_1에 속한 개체 $\boldsymbol{x}$를 π_2, π_3, $\cdots$, π_g로 잘못 분류할 기대 비용

잘못 분류될 총 기대 비용 :

$$ECM = p_1 ECM(1) + p_2 ECM(2) + \cdots + p_g ECM(g)$$

$$= \sum_{i=1}^{g} p_i [\sum_{k \neq i}^{g} P(k|i)c(k|i)]$$

정리 11.4 minimum *ECM* rule (Expected Cost of Misclassification)

개체 $\boldsymbol{x}$를 $\sum_{i \neq k}^{g} p_i f_i(\boldsymbol{x}) c(k|i)$ 를 최소로 하는 π_k로 분류

잘못 분류될 비용 $c(k|i)$들이 동일한 경우에는 잘못 분류될 확률을 최소화

: **minimum *TPM* rule (Total Probability of Misclassification)**

$$p_k f_k(\boldsymbol{x}) > p_i f_i(\boldsymbol{x}), \quad \text{모든 } i \neq k$$

또는

$$\ln p_k f_k(\boldsymbol{x}) > \ln p_i f_i(\boldsymbol{x}), \quad \text{모든 } i \neq k$$

이면, 개체 $\boldsymbol{x}$를 π_k로 분류

$\because$ $\sum_{i \neq k}^{g} p_i f_i(\boldsymbol{x})$를 최소로 하는 것은 $p_k f_k(\boldsymbol{x})$를 최대로 하는 것과 동일

2) 다변량 정규분포를 따르는 경우

$X_i \sim N_p(\mu_i, \Sigma_i)$: 정규모집단을 가정

p_i : 모집단 π_i 의 사전확률, $i = 1, 2, \ldots, g$

$$\ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{p}{2} ln(2\pi) - \frac{1}{2} ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)$$

를 최대로 하는 π_k로 관측값 $\boldsymbol{x}$를 분류

$\Leftrightarrow$ $\frac{p}{2} ln(2\pi)$는 모집단에 관계없이 동일하므로, 모집단 π_i의 일반화된 거리제곱 (generalized squared distance) 또는 **이차판별함수** (한 개체$\boldsymbol{x}$로부터 각 모집단의 중심까지의 거리)를 다음과 같이 정의하면

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2} ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln p_i$$

$$p_i f_i(\boldsymbol{x}) = \ln p_i - \frac{p}{2} ln(2\pi) - \frac{1}{2} ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)$$

$$= d_i^Q(\boldsymbol{x}) - \frac{p}{2} ln(2\pi)$$

이므로 $\ln p_k f_k(\boldsymbol{x})$를 최대로 하는 π_k로 $\boldsymbol{x}$를 분류하는 방법은

$d_k^Q(\boldsymbol{x}) = \max(d_1^Q(\boldsymbol{x}),\cdots,d_g^Q(\boldsymbol{x}))$이면 $\boldsymbol{x}$를 π_k로 분류하는 방법과 동일

① Σ_i들이 동일한 경우에는 선형판별함수를 사용

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i$$

i 번째 모집단에 대한 표본평균과 표본공분산행렬 : $\bar{\boldsymbol{x}}_i$, $\boldsymbol{S}_i$

전체 표본에 대한 합동공분산행렬 : $\boldsymbol{S}_p$

$$\boldsymbol{S}_p = \frac{(n_1-1)\boldsymbol{S}_1 + (n_2-1)\boldsymbol{S}_2 + \cdots + (n_g-1)\boldsymbol{S}_g}{n_1+n_2+\cdots+n_g-g}$$

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i'\boldsymbol{S}_p^{-1}\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}_p^{-1}\bar{\boldsymbol{x}}_i + \ln p_i$$

$\hat{d}_k(\boldsymbol{x}) = \max(\hat{d}_1(\boldsymbol{x}),\cdots,\hat{d}_g(\boldsymbol{x}))$이면 $\boldsymbol{x}$를 π_k로 분류

② 사전확률을 모를 경우

일반적으로 $p_1 = \ldots = p_g = 1/g$ 을 사용

③ Σ_i들이 동일한 경우에는 minimum TPM rule 은 사후확률을 최대로 하는 모집단을 선택하는 rule 과 동일

일반화된 거리 제곱을 이용하여 한 개체 $\boldsymbol{x}$가 k 번째 모집단에 속할 사후확률(posterior probability)

$$P(\pi_k|\boldsymbol{x}) = \frac{p_k f_k(\boldsymbol{x})}{\sum_{i=1}^{g} p_i f_i(\boldsymbol{x})} = \frac{\exp[-\frac{1}{2}d_k^Q(\boldsymbol{x})]}{\sum_{l}^{g}\exp[-\frac{1}{2}d_l^Q(\boldsymbol{x})]}$$

를 최대로 하는 π_k로 관측값 $\boldsymbol{x}$를 분류하는 방법과 동일

또는 다음과 같이 일반화된 거리를 정의하고

$$D_i^2(\boldsymbol{x}) = (\boldsymbol{x}-\bar{\boldsymbol{x}}_i)'\boldsymbol{S}_p^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)$$

$-\frac{1}{2}D_k^2(\boldsymbol{x}) + \ln p_k$ 가 최대인 π_k로 관측값 $\boldsymbol{x}$를 분류하는 방법과 동일

참고 : 정규가정에 매우 민감

Ex 11.11 (p614) Classifying a potential business-school graduate student

11.3 분류함수의 평가기준

(1) 두 모집단의 경우

$$TPM = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

을 최소로 하는 R_1과 R_2

$$R_1 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geqq \frac{p_2}{p_1}, \quad R_2 : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \leqq \frac{p_2}{p_1}$$

최소 TPM 분류함수의 오분류율(misclassification probability 또는 error rate)

$$Optimum\ \ Error\ \ Rate(OER) = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

표본분류함수의 평가 : 실제 오분류율

$$Actual\ \ Error\ \ Rate = AER = p_1 \int_{\hat{R}_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{\hat{R}_1} f_2(\boldsymbol{x}) d\boldsymbol{x},$$

단, $\hat{R}_1$과 $\hat{R}_2$는 각각 크기 n_1과 n_2인 표본을 이용하여 구한 범위로 $f_1(\boldsymbol{x})$과 $f_2(\boldsymbol{x})$에 의존

$\Rightarrow$ 따라서 $f_1(\boldsymbol{x})$과 $f_2(\boldsymbol{x})$에 의존하지 않는 비율을 구하기 위해 training 표본 중에서 잘못 분류된 표본비율을 이용

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

Confusion Matrix

소속모집단 \ 분류결과		π_1	π_2
	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}

n_{1C} : π_1에 속하는 표본중에서 π_1으로 옳게 분류된 표본의 개수

n_{1M} : π_1에 속하는 표본중에서 π_2 로 잘못 분류된 표본의 개수

n_{2C} : π_2에 속하는 표본중에서 π_2 로 옳게 분류된 표본의 개수

n_{2M} : π_2에 속하는 표본중에서 π_1으로 잘못 분류된 표본의 개수

$APER$의 문제점 : 분류함수의 추정에 사용된 표본들이 분류함수의 평가에도 이용되므로 오분류율을 과소 추정하는 경향이 있음

해결책

1) training sample과 validation sample

전체 표본을 training 표본과 validation 표본으로 나누어 training 표본은 분류함수의 추정에만 사용하고 평가에는 validation 표본만 사용함.

문제점 : 표본의 크기가 커야하며 또한 추정된 분류함수가 모든 표본을 사용한 것이 아니므로 실제로 구하고자 하는 분류함수가 아니라는 문제점이 있음

2) Lachenbruch's의 holdout procedure (Jackknife 또는 Cross-validation)

① π_1에서 한 개의 관측값을 유보(holdout)하고 나머지 n_1-1개의 표본과 π_2의 n_2개의 표본을 이용하여 분류함수를 구한다.

② ①에서 유보해 놓았던 표본을 추정된 분류함수를 이용하여 분류한다.

③ ①과 ②를 π_1에 속한 모든 관측값에 대해 반복하여 잘못 분류된 관측값의 개수 $n_{1M}^{(H)}$를 구한다.

④ ①-③을 π_2에 대해 시행하여 잘못 분류된 관측값의 개수 $n_{2M}^{(H)}$를 구한다.

$\Rightarrow$ Expected Actual Error Rate :

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2}$$

Ex. 11.8 (p 603) Classifying Alaskan and Canadian salmon

(2) $g(\geq 2)$ 모집단의 경우

holdout procedure에 의한 분류함수의 평가

$$\hat{E}(AER) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum_{i=1}^{g} n_i}$$

EX. 11.12 (p 619) Effective classification with fewer variables

11.4 Fisher의 판별식, $g(>2)$

모집단들을 가능한 한 분리하는 것이 목적

정규모집단의 가정은 필요하지 않음

공분산 행렬이 동일하다는 가정은 필요

$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$

시각적인 면과 그래픽을 이용한 분석에 중점

① 관측값들의 선형결합들을 이용한 차원의 축소를 통해 시각적으로 모집단들을 구분

② 중요한 처음 2~3 개의 선형결합들의 평균의 산점도를 이용하여 모집단들 사이의 관계 또는 집단화 정도를 파악

③ 처음 두 개의 판별함수 값의 산점도를 이용하여 이상점 또는 문제점을 파악

$\bar{\boldsymbol{\mu}} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu_i}$: 모집단의 총 평균벡터

$\boldsymbol{B_\mu} = \sum_{i=1}^{g}(\boldsymbol{\mu_i}-\bar{\boldsymbol{\mu}})(\boldsymbol{\mu_i}-\bar{\boldsymbol{\mu}})'$: 모집단 간의 교차제곱합 행렬

$Y = \boldsymbol{\ell}'\boldsymbol{x}$: 선형결합

$\mu_{iY} = E(Y) = \boldsymbol{\ell}'E(\boldsymbol{X}|\pi_i) = \boldsymbol{\ell}'\boldsymbol{\mu_i}, \quad i = 1,2,\ ...,g$

$Var(Y) = \boldsymbol{\ell}'Cov(\boldsymbol{X})\boldsymbol{\ell} = \boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}$: Y의 분산

$\bar{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g}\mu_{iY} = \frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\ell}'\boldsymbol{\mu_i} = \boldsymbol{\ell}'(\frac{1}{g}\sum_{i=1}^{g}\boldsymbol{\mu_i}) = \boldsymbol{\ell}'\bar{\boldsymbol{\mu}}$: 총평균

$$\frac{\text{각 그룹평균과 총평균간의 거리의 제곱합}}{Y\text{의 분산}} = \frac{\sum_{i=1}^{g}(\mu_{iY}-\bar{\mu}_Y)^2}{\sigma_y^2} = \frac{(\boldsymbol{\ell}'\boldsymbol{\mu_i}-\boldsymbol{\ell}'\bar{\boldsymbol{\mu}})^2}{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}} = \frac{\boldsymbol{\ell}'\boldsymbol{B_\mu}\boldsymbol{\ell}}{\boldsymbol{\ell}'\boldsymbol{\Sigma}\boldsymbol{\ell}}$$

: **그룹 내의 분산과 그룹 간의 분산의 비**

그룹 내의 분산과 그룹 간의 분산의 비가 최대가 되도록 $\boldsymbol{\ell}$을 선택

$\bar{\boldsymbol{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}$: 표본 평균벡터

$\bar{\boldsymbol{x}} = \frac{1}{g}\sum_{i=1}^{g}\bar{x}_i$: 표본 총평균 벡터

$\hat{\boldsymbol{B}} = \sum_{i=1}^{g}(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})'$: 표본 그룹간 교차제곱합 행렬

$\boldsymbol{W} = \sum_{i=1}^{g}(n_i-1)\boldsymbol{S}_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)'$: 표본 그룹내 교차제곱합 행렬

$\Rightarrow \boldsymbol{S_p} = \boldsymbol{W}/(n_1+n_2+\cdots+n_g) = \hat{\boldsymbol{\Sigma}}$

정리 : Fisher의 표본판별식

$\hat{\lambda}_1,\cdots,\hat{\lambda}_s > 0$: $\boldsymbol{W}^{-1}\hat{\boldsymbol{B}}$의 고유값, $s \leqq \min(g-1,p)$

$\hat{\boldsymbol{e}}_1,\hat{\boldsymbol{e}}_2,\cdots,\hat{\boldsymbol{e}}_s$: 고유값 ($\hat{\boldsymbol{e}}'\boldsymbol{S}_p\hat{\boldsymbol{e}} = 1$을 만족) 이라고 하면 $\hat{\boldsymbol{\ell}}_1 = \hat{\boldsymbol{e}}_1$은

$$\frac{\hat{\boldsymbol{\ell}}'\hat{\boldsymbol{B}}\hat{\boldsymbol{\ell}}}{\hat{\boldsymbol{\ell}}'\boldsymbol{W}\hat{\boldsymbol{\ell}}} = \frac{\hat{\boldsymbol{\ell}}'\left[\sum_{i=1}^{g}(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})'\right]\hat{\boldsymbol{\ell}}}{\hat{\boldsymbol{\ell}}'\left[\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)'\right]\hat{\boldsymbol{\ell}}}$$

를 최대로 하는 선형결합이며,

$\hat{\boldsymbol{\ell}}_1'\boldsymbol{x} = \hat{\boldsymbol{e}}_1'\boldsymbol{x}$: **Fisher의 제 1 표본판별식**

$\hat{\boldsymbol{\ell}}_2'\boldsymbol{x} = \hat{\boldsymbol{e}}_2'\boldsymbol{x}$: **Fisher의 제 2 표본판별식**

$\hat{\boldsymbol{\ell}}_k'\boldsymbol{x} = \hat{\boldsymbol{e}}_k'\boldsymbol{x}$: **Fisher의 제 k 표본판별식,** $k \leqq s$.

Ex. 11.14 (Fisher's discriminants for the crude-oil data)

분류(Classification)

1) Fisher의 discriminants 이용

$$\sum_{j=1}^{r}(\hat{y}_j-\bar{y}_{kj})^2=\sum_{i=1}^{r}[\hat{\boldsymbol{\ell}}_j'(\boldsymbol{x}-\bar{\boldsymbol{x}}_k)]^2 \leqq \sum_{j=1}^{r}[\hat{\ell}_j(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)]^2 \text{ , } i\neq k \text{ 이면 } \boldsymbol{x}\text{를 } \pi_k\text{로 분류}$$

2) 로지스틱 회귀 방법(logistic regression approach)

분류함수를 구하는데 사용될 변수들이 질적변수 또는 범주형변수인 경우(특히 0-1 값을 갖는 변수)에는 로지스틱회귀 방법을 사용하는 것이 적합

$$logit(p)=\ln(odds)=\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 z$$

$$\theta(z)=\frac{p(z)}{1-p(z)}=\exp(\beta_0+\beta_1 z)$$

$$p(z)=\frac{\exp(\beta_0+\beta_1 z)}{1+\exp(\beta_0+\beta_1 z)} \quad : \text{logistic curve}$$

logistic regression

$$\ln\left(\frac{p(\boldsymbol{z})}{1-p(\boldsymbol{z})}\right)=\beta_0+\beta_1 z_1+\cdots+\beta_r z_r=\boldsymbol{\beta}'\boldsymbol{z_j},\ \boldsymbol{\beta}=[\beta_0,\beta_1,\ \ldots,\beta_r]'$$

분류법

$$\ln\frac{\hat{p}(\boldsymbol{z})}{1-\hat{p}(\boldsymbol{z})}=\hat{\beta}_0+\hat{\beta}_1 z_1+\cdots+\hat{\beta}_r z_r=\hat{\boldsymbol{\beta}}'\boldsymbol{z_j} \ > 0 \text{이면 } \boldsymbol{z}\text{를 모집단 1로 분류}$$

Ex 11.17 (p639) Logistic regression with the salmon data

3) Classification Trees

Classification and Regression Trees(CART)

모집단의 분포 또는 최적의 기준과는 무관

4) Neural Networks

5) 변수 선택(Stepwise Discriminant Analysis)

변수의 수가 많은 경우 대부분의 정보를 간직하고 있는 적은 개수의 변수를 선택하여 분류함수를 구하는 것이 바람직

최적이라는 보장이 없음

6) 그룹평균의 동일성 여부 검정(MANOVA 이용)

모집단들이 잘 분리가 되어있지 않은 경우에는 분류가 잘 안됨.

MANOVA를 이용하여 그룹 평균들이 동일한지 여부를 사전에 검정

7) 그래픽을 이용한 방법

8) 정규성가정이 어려운 경우

처음 몇 개의 Fisher의 선형판별식들을 이용하여 정규성 검정 가능

11.4 정준판별분석

(1) 두 모집단의 경우

정준판별분석은 정준변수(canonical variable)를 이용한 판별분석

첫 번째 모집단을 나타내는 변수 : $\boldsymbol{x_1}' = (x_{11}, x_{12}, \cdots, x_{1p})$

두 번째 모집단을 나타내는 변수 : $\boldsymbol{x_2}' = (x_{21}, x_{22}, \cdots, x_{2p})$

두 모집단 사이의 상관관계를 최대로 하는 변수들의 선형결합

$$y_1 = \ell_{11}x_{11} + \ell_{12}x_{12} + \cdots + \ell_{1p}x_{1p}$$

$$y_2 = \ell_{21}x_{21} + \ell_{22}x_{22} + \cdots + \ell_{2p}x_{2p}$$

y_1, y_2의 상관계수가 최대가 되도록 하는 계수를 구한다.

$\boldsymbol{\ell_1}' = (\ell_{11}, \ell_{12}, \cdots, \ell_{1p})$와 $\boldsymbol{\ell_2}' = (\ell_{21}, \ell_{22}, \cdots, \ell_{2p})$

정준상관계수(canonical correlation coefficient) : y_1과 y_2의 최대상관계수

(2) g개의 모집단을 판별하는 문제

$\boldsymbol{x}_1$ $p \times 1$: 관측값 변수

$\boldsymbol{x}_2$ $(g-1) \times 1$: 모집단 또는 그룹을 나타내는 가상적인 변수

$\boldsymbol{x}_2$의 관측값은 i번째 모집단에서 관측된 데이터에 대하여 i번째 원소만 1이고 나머지는 모두 0인 단위벡터로 간주

$\boldsymbol{x}_1$과 $\boldsymbol{x}_2$의 관측값을 이용하여 $\boldsymbol{x}_1$ 만의 정준변수를 고려

$\boldsymbol{x}_1$의 제1 정준변수 : $\boldsymbol{x}_2$의 선형결합과 제일 강한 상관관계를 가지므로, 이 제1 정준변수는 모집단의 차이를 가장 잘 설명하는 변수

마찬가지로 $\boldsymbol{x}_1$의 제2, 제3, …, 제$g-1$ 정준변수를 구하면 이들은 서로 무상관이면서 모집단의 차이를 가장 잘 설명하는 변수가 된다.

처음 2개 또는 3개에 의해 관측값을 나타내면 모집단 또는 그룹의 판별이 가능

정준변수를 흔히 '판별함수'라고 부르기도 하나, 이의 개념은 DISCRIM 절차에서 다룬 판별함수와는 개념이 많이 다르므로, 여기서는 정준변수라는 용어를 쓰도록 한다.

CANDISC 절차에 의한 판별함수는 정준변수를 원하는 개수만큼으로 줄여서, 이들을 이용하여 주어진 자료를 분석하는 것이다. 이 때에 정준변수의 수는 대응되는 고유값의 크기에 따라 결정된다. 즉, 요인분석에서 인자의 개수를 정하는 것과 비슷한 개념이다.

11.5 SAS에서 제공되는 판별분석을 위한 절차

① DISCRIM 절차

g 개의 모집단에서 관측된 기존의 (어느 모집단에 속하는지를 알고 있는) 관측값에 기초하여 판별함수를 만들고 이를 기초로 새로운 관측값을 분류한다. 정규성을 가정한 모수적 방법과 분포에 대한 가정을 하지 않는 비모수적 방법이 모두 가능하다.

② CANDISC 절차

정준판별분석을 실시하는 절차로서, 주성분분석 및 정준상관(canonical correlation)과 매우 밀접한 관계를 갖고 있다. 즉, 주어진 변수들의 일차결합으로 정준변수(canonical variable)를 만들어 차원축소(dimension reduction)를 시행하고, 이를 판별함수로 사용한다. 주성분은 전체 변동을 가장 설명하는 순서대로 제1주성분, 제2주성분, … 이 결정되지만, 정준변수는 g 개의 모집단간의 차이를 가장 잘 설명하는 순서대로 제1정준, 제2정준, … 이 결정된다.

③ STEPDISC 절차

STEPDISC 절차는 stepwise 회귀분석과 비슷한 개념으로 이해할 수 있다. 즉, VAR로 선언한 변수들 가운데 판별에 유용한 변수를 선택한다. 선택방법은 변수추가법(forward selection), 변수제거법(backward elimination), 변수증감법(stepwise selection)이 모두 가능하다. 모집단 간의 차이를 잘 나타내는 변수를 단계적(stepwise)으로 선택하여 이들을 판별함수에 사용한다.

(1) DISCRIM 절차의 이용

1) DISCRIM 절차의 일반형

```
PROC DISCRIM <options>;
  CLASS variable;
  VAR variables;
  PRIORS probabilities;
  FREQ variables;
  WEIGHT variable;
  ID variable;
  TESTFREQ variable;
  TESTID variable;
  BY variables;
```

DISCRIM 절차에서 사용되는 SAS 문과 옵션

① PROC DISCRIM <options>;

i. Data Set options

DATA=SAS dsn(data set name) : 분석에 사용될 자료명. 사용될 수 있는 종류는 원래의 관측값이거나 TYPE=CORR, COV, CSSP, SSCP, LINEAR, QUAD, MIXED와 같이 SAS/STAT 절차에 의해 만들어진 자료명. METHOD=NPAR이 사용되면 반드시 원래의 관측값으로 이루어진 자료명이 사용되어야 한다.

TESTDATA=SAS dsn : 분류함수에 의해 분류하고자 하는 관측값들이 기억될 자료명을

지정

OUTSTAT=SAS dsn : 평균, 표준편차, 상관계수등과 같은 분석결과의 통계량의 값이 기억될 자료명을 지정. 이상의 통계량 외에도 METHOD=NORMAL이 사용되면 판별함수의 계수가 기억되며 출력 자료명은 TYPE=LINEAR(POOL=YES), QUAD(POOL=NO) 또는 MIXED(POOL=TEST) 중 하나. METHOD=NPAR이 사용되면 TYPE=CORR

OUT=SAS dsn(또는 **OUTCROSS=SAS dsn**) : 관측값, 사후확률과 resubstitution 방법(또는 crossvalidation 방법)에 의해 분류된 결과가 기억될 자료명을 지정. 비모수적 방법의 경우는 사용할 수 없다.

OUTD=SAS dsn : 관측값과 추정된 밀도함수가 기억될 자료명을 지정

TESTOUT=SAS dsn : 'TESTDATA='에 저장된 관측값, 사후확률과 분류결과가 기억될 자료명을 지정

ii. 판별함수의 종류를 선택하는 옵션들

METHOD=NORMAL(또는 **METHOD=NPAR**) : 모수적방법(또는 비모수적방법)에 의해 판별함수를 구한다.

POOL=YES : 합동공분산(pooled covariance)추정량을 이용

=NO : 그룹내 공분산(within-group-covariance) 추정량을 이용

=TEST : 그룹내 공분산의 동일성 검정을 시행한다. 검정결과가 'SLPOOL='에 의해 지정된 수준에서 유의하면 그룹내 공분산추정량을 이용하고, 유의하지 않으면 합동공분산추정량을 이용

SLPOOL=p : 동일성 검정의 유의수준을 지정. 디폴트(default)의 경우에는 .10이 사용.

iii. Resubstitution방법(또는 crossvalidation방법)과 관련된 옵션들

LIST(또는 **CROSSLIST**) : 모든 관측값에 대한 resubstitution 방법(또는 crossvalidation 방법)에 의한 분류결과를 출력

ILSTERR(또는 **CROSSLISTERR**) : resubstitution 방법(또는 crossvalidation 방법)에 의해 잘못 분류된 관측값만을 출력

NOCLASSIFY : resubstitution 방법에 의한 분류결과를 출력하지 않음

CROSSVALIDATE : crossvalidation 방법에 의해 분류

iv. Test data의 분류에 관한 옵션들

TESTLIST : 'TESTDATA='에 저장된 관측값들의 분류결과를 출력

TESTLISTERR : 'TESTDATA='에 저장된 관측값들 중 잘못 분류된 관측값만을 출력한다. TESTCLASS 문이 사용된 경우에만 사용

v. 출력과 관련된 옵션과 기타 옵션들

SINGULAR=p : 표본공분산 행렬의 singularity를 결정하는 기준을 지정, 0<p<1. 디폴트인 경우는 p=1E-8이다.

BCORR : 그룹간의 표본상관행렬을 출력

PCORR : 합동그룹내 표본상관행렬을 출력

ALL : 모든 결과를 출력

ANOVA : 각 변수별로 모든 그룹의 모평균이 동일하다는 가설의 검정을 위한 일변량 분산분석 결과를 출력

MANOVA : 모든 그룹의 모평균벡터들이 동일하다는 가설의 검정을 위한 다변량 분산분석 결과를 출력.

DISTANCE : 그룹평균간의 거리를 출력

SIMPLE : 전체표본과 각 모집단별로 각 변수들에 대한 단순기술통계량을 출력

NOPRINT : 출력하지 않는다.

SHORT : 'METHOD=NORMAL'이 사용되면 그룹평균간의 거리와 판별함수의 계수를 출력하지 않는다.

② CLASS variable;

분류변수로 사용될 변수를 지정. DISCRIM 절차에서 반드시 사용되어야 한다.

③ VAR variables;

분석에 사용될 양적변수들을 지정. 디폴트의 경우에는 모든 숫자변수를 사용

④ ID variable;

LIST 또는 LISTERR 문이 사용된 경우에만 유용하게 쓰이며, 분류결과를 프린트할 때 각 관측값 대신에 이에 해당하는 ID 변수가 출력

⑤ PRIOR probabilities;

분류시에 사용할 사전확률을 지정. 디폴트이면 동일사전확률이 사용. 사전확률을 표본의 크기에 비례하도록 하고 싶을 때는

```
PRIORS proportional;
```

을 사용

⑥ FREQ variable;

SAS dsn에 있는 어떤 변수의 값이 해당되는 값이 포함된 관측값의 관측도수를 의미할 경우, 이 변수를 FREQ 문에 명시하면 해당관측값이 변수의 값에 해당되는 회수만큼 반복 관측된 것으로 생각하고 처리. 자유도 및 관측값의 개수를 계산할 때 반영

⑦ TESTID variable;

이 문이 사용되면 TESTLIST 문과 TESTLISTERR 문이 역시 사용되며 새로 분류하고자 하는 TESTDATA의 분류결과를 프린트할 때 관측값 대신에 프린트될 변수를 지정. 이 문에서 저정된 변수들은 반드시 TESTDATA=SAS dsn으로 지정된 SAS dsn에 있어야 한다.

⑧ TESTFREQ variable;

FREQ 문과 마찬가지로 'TESTDATA='에 저장된 관측값의 관측도수를 의미하는 변수를 지정할 경우 사용

참고 결측값(Missing Value)의 처리

분류변수가 결측인 관측값들은 분류함수를 구할 때는 사용되지 않으나 분류함수에 의해 분류는 된다. 다른 변수값들이 결측이면 분석에서 제외된다.

2) 예제

예제 11.1(계속) 운전용 잔디깍는 기계의 판매전략을 위하여 판별함수를 구하고, 두 관측값에 대하여 DISCRIM 절차를 이용한 분류를 시행

```
/*  DISCRIM1.SAS  : DISCRIMINANT ANALYSIS OF LAWN MOWER DATA  */
TITLE 'DISCRIMINANT ANALYSIS FOR LAWN MOWER';
OPTIONS PS=60 PAGENO=1;
DATA LAWN;
  INPUT OWN $ INCOME SIZE  @@;
  CARDS;
O 60.0 18.4   O  85.5 16.8   O  64.8 21.6   O 61.5 20.8
O 87.0 23.6   O 110.1 19.2   O 108.0 17.6   O 82.8 22.4
O 69.0 20.0   O  93.0 20.8   O  51.0 22.0   O 81.0 20.0
N 75.0 19.6   N  52.8 20.8   N  64.8 17.2   N 43.2 20.4
N 84.0 17.6   N  49.2 17.6   N  59.4 16.0   N 66.0 18.4
N 47.4 16.4   N  33.0 18.8   N  51.0 14.0   N 63.0 14.8
;
PROC PLOT;
  PLOT SIZE*INCOME=OWN / VPOS=35 VAXIS=12 TO 24 BY 3
                         HPOS=70 HAXIS=30 TO 120 BY 30;
RUN;

PROC DISCRIM DATA=LAWN  POOL=TEST SIMPLE LISTERR OUT=LACAL; ① ② ③
  CLASS OWN;
  VAR INCOME SIZE ;
  PRIORS PROP; ④
PROC PRINT DATA=LACAL;
DATA TEST; ⑤
  INPUT OWN $ INCOME SIZE XVALUES $4-12;
  CARDS;
.  60.5 19.5
.  79.0 18.0
;
PROC DISCRIM DATA=LACAL TESTDATA=TEST TESTLIST; ⑥
  CLASS OWN;
  TESTID XVALUES;
  VAR INCOME SIZE;
  TITLE2 'TEST OF CLASSIFICATION FUNCTION';
RUN;
```

예제12.2(계속) 경영대학원의 입학허가 여부를 결정하는 판별분석

```
/* DISCRIM2.SAS  :  DISCRIMINANT ANALYSIS OF ADMISSION DATA */
DATA ADMIT;
  INPUT GROUP GPA GMAT @@;
  CARDS;
1 2.96 596 2 2.54 446 3 2.86 494 1 3.14 473 2 2.43 425 3 2.85 496
1 3.22 482 2 2.20 474 3 3.14 419 1 3.29 527 2 2.36 531 3 3.28 371
1 3.69 505 2 2.57 542 3 2.89 447 1 3.46 693 2 2.35 406 3 3.15 313
1 3.03 626 2 2.51 412 3 3.50 402 1 3.19 663 2 2.51 458 3 2.89 485
1 3.63 447 2 2.36 399 3 2.80 444 1 3.59 588 2 2.36 482 3 3.13 416
1 3.30 563 2 2.66 420 3 3.01 471 1 3.40 553 2 2.68 414 3 2.79 490
1 3.50 572 2 2.48 533 3 2.89 431 1 3.78 591 2 2.46 509 3 2.91 446
```

```
1 3.44 692 2 2.63 504 3 2.75 546 1 3.48 528 2 2.44 336 3 2.73 467
1 3.47 552 2 2.13 408 3 3.12 463 1 3.35 520 2 2.41 469 3 3.08 440
1 3.39 543 2 2.55 538 3 3.03 419 1 3.28 523 2 2.31 505 3 3.00 509
1 3.21 530 2 2.41 489 3 3.03 438 1 3.58 564 2 2.19 411 3 3.05 399
1 3.33 565 2 2.35 321 3 2.85 483 1 3.40 431 2 2.60 394 3 3.01 453
1 3.38 605 2 2.55 528 3 3.03 414 1 3.26 664 2 2.72 399 3 3.04 446
1 3.60 609 2 2.85 381 1 3.37 559 2 2.90 384 1 3.80 521 1 3.76 646
1 3.24 467
;
PROC PLOT;
  PLOT GMAT*GPA=GROUP / VPOS=30; RUN;
PROC DISCRIM DATA=ADMIT POOL=TEST SLPOOL=0.05  LISTERR OUT=ADMI; ① ② ③
  CLASS GROUP;
  VAR GPA GMAT; RUN;
DATA TEST;
  INPUT GROUP GPA GMAT;
  CARDS;
 .  3.21  497
;
PROC DISCRIM DATA=ADMI TESTDATA=TEST TESTLIST; ④
  CLASS GROUP;
  VAR GPA GMAT; RUN;
```

(2) CANDISC 절차의 이용

1) CANDISC 절차의 일반형

```
PROC CANDISC <options>;
  VAR variables;
  CLASS variable;
  FREQ variable;
```

(대부분의 옵션들은 DISCRIM과 같음)

CANDISC 절차에서 사용되는 SAS 문과 옵션들은 다음과 같다.

NCAN=n : 정준변수의 개수를 지정한다. 디폴트는 변수의 개수와 같다.

PREFIX=name : 정준변수의 이름을 지정한다. 디폴트는 CAN1, CAN2, … 등이다.

이 외에 **OUT= SAS dsn, SIMPLE, ANOVA, ALL, WCOV, TCOV**, … 등의 많은 옵션들이 있다.

2) 예제

예제 11.3 (Fisher의 붓꽃자료 분석) 붓꽃의 세 품종(setosa, versicolor, virginica)의 각각에서 50개씩 추출하여 다음의 4변수를 mm단위로 측정(SAS/STAT 메뉴얼).

x_1 : sepal length (꽃받침의 길이); x_2 : sepal width (꽃받침의 폭)

x_3 : petal length (꽃잎의 길이); x_4 : petal width (꽃잎의 폭)

표에서 $(x_1, x_2, x_3, x_4,$품종번호$)$의 관측값을 연속적으로 나열하였으며, 품종번호는

1=setosa, 2=versicolor, 3=virginica

를 뜻한다. 이들 자료에 대하여 CANDISC 절차를 이용한 판별분석을 시행해보자.

FORMAT 절차는 versicolor와 virginica의 첫 글자가 같기 때문에 산점도에서 이들을 구별하는 기호를 지정해준 것이다.

```
/* CANDISC.SAS : ANALYSIS OF FISHER'S IRIS DATA */
OPTIONS PS=60 PAGENO=1;
PROC FORMAT;
  VALUE SPECNAME
     1='SETOSA     '
     2='VERSICOLOR'
     3='VIRGINICA ';
  VALUE SPECCHAR
     1='S'
     2='O'
     3='V';
RUN;
DATA IRIS;
  INFILE 'B:\IRIS.DAT';
  INPUT X1-X4 SPECIES @@;
  FORMAT SPECIES SPECNAME.;
  LABEL X1='SEPAL LENGTH'
        X2='SEPAL WIDTH '
        X3='PETAL LENGTH'
        X4='PETAL WIDTH ';
PROC CANDISC DATA=IRIS OUT=CANOUT; ① ②
  CLASS SPECIES;
  VAR X1-X4;
RUN;
PROC PRINT DATA=CANOUT;
RUN;
PROC PLOT; ③
  PLOT CAN2*CAN1=SPECIES / VPOS=35;
  FORMAT SPECIES SPECCHAR.;
RUN;
```

(3) STEPDISC 절차의 이용

1) STEPDISC 절차의 일반형

```
PROC STEPDISC <options>;
  VAR variables;
  CLASS variable;
```

이 외에 FREQ, WEIGHT, BY 등이 있다.

STEPDISC 절차에서 사용되는 SAS 문과 옵션들은 다음과 같다.

METHOD=FW | FORWARD : 변수추가법으로 지정
=BW | BACKWARD : 변수제거법으로 지정
=SW | STEPWISE : 변수증감법으로 지정
(디폴트는 STEPWISE이다.)

SLE=P : 변수추가법에서 p-값 지정(디폴트는 0.15)
SLS=P : 변수제거법에서 p-값 지정(디폴트는 0.15)

2) 예제

예제 11.3(계속) 피셔의 붓꽃자료를 STEPDISC 절차로 분석

STEPWISE 방법을 이용하는 SAS 프로그램

```
/* STEPDISC.SAS : ANALYSIS OF FISHER'S IRIS DATA */
OPTIONS PS=60 PAGENO=1;
PROC FORMAT;
  VALUE SPECNAME
     1='SETOSA     '
     2='VERSICOLOR'
     3='VIRGINICA ';
  VALUE SPECCHAR
     1='S'
     2='O'
     3='V';
RUN;
DATA IRIS;
  INFILE 'B:\IRIS.DAT';
  INPUT X1-X4 SPECIES @@;
  FORMAT SPECIES SPECNAME.;
  LABEL X1='SEPAL LENGTH'
        X2='SEPAL WIDTH '
        X3='PETAL LENGTH'
        X4='PETAL WIDTH ';
RUN;
PROC STEPDISC DATA=IRIS METHOD=STEPWISE;
  CLASS SPECIES;
  VAR X1-X4;
RUN;
```