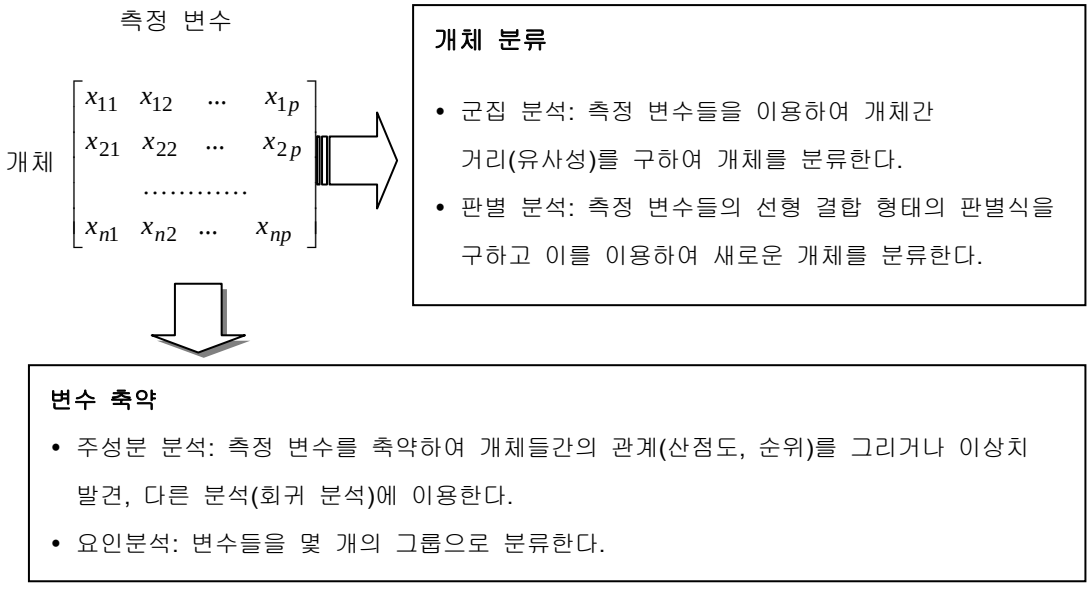
CHAPTER 6

판별 분석

6.1. 개요

주성분 분석과 요인 분석은 $x_1, x_2, ..., x_p$ 변수들의 상관 관계(공분산 행렬이나 상관계수 행렬)를 이용하여 변수의 차원을 축약하거나 변수를 그룹화하는 방법으로 변수 유도 기법(Variable-directed techniques)이라 한다. 판별 분석(Discriminant Analysis)은 군집 분석(Clustering Analysis)과 함께 개체들에 대해 측정된 특성(변수) 값을 이용하여 개체를 분류하는 방법으로 개체 유도 기법(individual directed techniques)라 일컫는다.

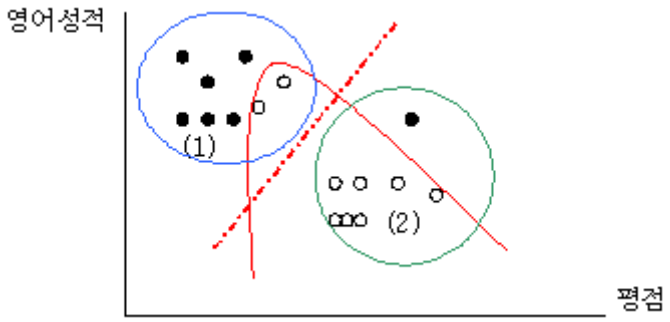


6.1.1. 군집 분석과 판별 분석

○○대학교 졸업생 40 명에 대해 평점, 비만도(=키/몸무게: 외모), 영어 성적, 자격증 개수, 원서 지원 회수, 가족 총 연 소득 (재정 능력), 친구 수(사교력 측정) 조사하였다고 하자. 측정된 7 개의 변수의 거리(distance) 혹은 유사성(similarity) 개념을 이용하여 개체를 분류하는 것을 군집 분석이라 한다. 개체 그룹의 수는 분석자가 임의로 정하게 되며 그룹 내의 개체를 본 후 그 그룹의 이름을 부여한다.

판별 분석은 개체의 그룹이 분류 전에 조사되어 있고 그룹 변수와 측정 변수들을 이용하여 개체의 그룹을 판별하는데 적절한 판별식을 구하고 이를 이용하여 새로운 개체를 분류한다. 만약 학생을 그룹(집단)으로 분류할 수 있는 취업 유무를 조사하였다면 7 개의 변수들을 이용하여 개체들을 나눌 수 있는 판별식을 구하고 이를 이용하여 새로운 학생들이 취업을 할 수 있을지 (취업 집단에 속하는지, 취업 확률) 알아내는 것을 판별 분석이라 한다. 이처럼 군집 분석과 판별 분석의 차이는 그룹이 분석 전에 알려져 있는가 아닌가 하는 것이다.

16 명(남자 9 명, 여자 7 명)에 대해 영어 성적과 평점만을 조사해 다음 산점도를 얻었다고 하자.



군집 분석에서는 남녀 구별이 없는(●, ○ 구분 없음) 상태에서 가까운 개체(유사성이 높다) 들끼리 묶어가는 것이다. 개체를 두 그룹으로 나눌 수 있고 (물론 3 개 이상의 그룹으로도 나눌 수 있다. 분석자가 집단 간의 거리를 보고 판단하게 된다) 원에 속한 개체끼리 묶을 수 있을 것이다. 각 그룹의 이름은 그룹에 속한 개체의 속성을 보고 붙인다. 까만 원에 속한 개체는 영어 성적이 높으므로 영어 성적 우위 그룹, 아래 파랑 원은 평점 우위 그룹으로 이름을 붙일 수 있을 것이다. 측정 변수가 2 개인 경우 산점도를 그리면 군집(개체 그룹)의 이름을 붙일 수 있으나 측정 변수가 3 개 이상인 경우에는 군집의 이름 부여가 어렵다. 그러므로 이런 경우 주성분 분석을 이용하여 변수를 축약하고 주성분 변수를 이용하여 군집

분석을 실시하면(산점도를 그릴 수 있다) 된다. 주성분 분석을 이용한 군집 분석은 주성분 분석 사용 예에서 중요하다.

판별 분석은 자료 수집 시 이미 그룹이 나누어져 있으므로(남:○, 여:●) 1)개체(사람)가 어느 남녀 그룹에 속하는지 판별하는 식을 구하고 2)이를 이용하여 새로운 개체를 분류하게 된다. (영어 성적과 평점을 알면 그 사람의 성별을 판별할 수 있다) (1)과 (2)는 개체를 분류하는 판별식의 예이다. (2)판별식이 개체 분류의 오류가 없으나 이런 곡선 식을 구하는 것은 거의 불가능하다. 그러므로 구하기 쉬운 직선 형태의 판별식(Fisher 의 Linear Deterministic Function 이라 한다) (1)을 이용하게 된다.

6.1.2. 오분류

마취과 의사는 심장 수술에 마취가 안전한지 알아보기 위하여 나이, 혈압, 몸무게 등을 조사하고 마취 후 안전 여부(그룹)를 조사하였다고 하자. 마취과 의사를 알고 싶다 (1)이 자료를 토대로 새로운 환자가 왔을 때 마취가 안전한지 판단할 수 있을까? 이런 판별 규칙을 판별식이라 한다. (2)이 판별식을 사용하였을 때 개체를 잘못 분류할 확률, 즉 오분류(misclassification) 확률은 얼마인가?

판별 / 실제	마취 안전	마취 위험
마취 안전	정분류	오분류①
마취 위험	오분류②	정분류

개체의 집단이 2 개인 경우 오분류는 2 가지 경우가 생긴다. 위의 예를 살펴보면 (1)마취를 해도 괜찮은 환자를 마취하면 안 되는 환자로 분류하거나(오분류①) (2)마취를 해서 안 되는 환자를 마취해도 괜찮은 환자로 분류하는 (오분류②) 잘못을 저지르게 된다. 이 예제에서 오분류①이 발생하면 환자에게 고통을 주거나 병원 수입이 줄어들게 되고 오분류②는 의료사고로 이어질 수 있으므로 오분류②에 의해 발생하는 비용이 훨씬 크다. 일반적으로 오분류 비용 계산이 어려우므로 오분류 비용은 동일하다고 가정한다. 그러므로 판별 분석은 오분류를 최소화 할 수 있는 판별식을 구하는 것이 주목적이다.

6.2. 모집단이 두 개인 경우

다변량 정규분포를 따르는 2 개의 모집단이 있다고 가정하자.

모집단 1: $\pi_1 \sim N_p(\underline{\mu}_1, \Sigma_1)$, 모집단 2: $\pi_2 \sim N_p(\underline{\mu}_2, \Sigma_2)$

각 모집단으로부터 n_1, n_2 개의 표본을 뽑아 각 개체에 대해 p 개 변수를 측정하였다고 하고 한 개체의 측정치를 $\underline{x}_0$ 라 하자. 그리고 오분류에 의한 비용 함수가 같다고 가정하자.

아래 모든 판별식(규칙)들은 모집단의 모수가 있으므로 이에 대한 추정치가 필요하므로 모평균($\underline{\mu}$)은 표본 평균($\bar{x}$)으로, 분산-공분산 행렬(Σ)은 표본 분산-공분산($\hat{\Sigma} = S$)을 사용한다.

통합(pooled) 분산-공분산 행렬은 $\hat{\Sigma} = \frac{(n_1-1)\hat{\Sigma}_1 + (n_2-1)\hat{\Sigma}_2}{(n_1+n_2-2)} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{(n_1+n_2-2)}$ 이다.

6.2.1. 판별 규칙

개체를 판별하는 규칙을 판별식이라 하며 다음과 같은 규칙이 있다.

(1)우도(Likelihood) 규칙

개체에 대해 $L(\underline{x}_0 : \underline{\mu}_1, \Sigma_1) > L(\underline{x}_0 : \underline{\mu}_2, \Sigma_2)$ 이면 π_1 으로 분류하고, $L(\underline{x}_0 : \underline{\mu}_1, \Sigma_1) < L(\underline{x}_0 : \underline{\mu}_2, \Sigma_2)$ 이면 π_2 로 분류한다.

$$L(\underline{x}_p : \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp^{[-1/2(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})]}$$

(2)선형 판별식(Fisher's Linear Discriminant Function) 규칙

두 모집단이 동일한 분산-공분산 행렬을 (variance-covariance Σ) 갖는다면 위의 우도 함수 규칙은 (likelihood function) 다음과 같이 간단화 된다. 만약 $\underline{b}' \underline{x}_0 - k > 0$ (Linear Discriminant function)이면 π_1 으로 분류하고, 그렇지 않으면 π_2 으로 분류한다.

$$\underline{b}' = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}, \quad k = (1/2)(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

(3)Mahalanobis 거리 규칙

모집단이 동일한 분산-공분산 행렬을 (variance-covariance Σ) 갖는다면, 우도 함수 규칙은 다음과 동일하다. 만약 $d_1 < d_2$ 이면 π_1 으로 분류하고, 그렇지 않으면 π_2 으로 분류한다.

$$\text{Mahalanobis Distance: } d_i = (\underline{x}_0 - \underline{\mu}_i)'\Sigma^{-1}(\underline{x}_0 - \underline{\mu}_i) \quad i = 1,2.$$

(4)사후 확률(Posterior Probability) 규칙

모집단이 동일한 분산-공분산 행렬을 가질 때 모집단 π_1 의 사후 확률을 다음과 같이 정의하고 만약 $P(\pi_1 \mid \underline{x}_0) > P(\pi_2 \mid \underline{x}_0)$ 이면 π_1 으로 분류하고, 그렇지 않으면 π_2 으로 분류한다.

$$P(\pi_i \mid \underline{x}_p) = \frac{\exp^{(-1/2)d_i}}{\exp^{[-1/2d_1]} + \exp^{[-1/2d_2]}}$$

6.2.2. 오분류 비율 추정

집단이 2 개인 경우 오분류는 1 집단에 속한 개체를 사용된 판별식에 의해 2 집단으로 분류하거나 2 집단 속한 개체를 1 집단으로 분류하는 경우이다. 오분류가 적은 판별식이 선호된다. 물론 비용 함수가 존재한다면 판별식 선택 시 비용까지도 고려해야 하지만 비용 계산은 쉽지 않고 비용 함수 설정은 다소 주관적일 가능성이 높다.

(1)Re-substitution 규칙

수집된 데이터로부터 얻은 판별식을 원 데이터에 적용하여 개체를 분류하여 오분류 비율을 구하는 것으로 정분류 비율이 높게 추정될(overestimate) 가능성이 있어 거의 사용하지 않는다.

(2)테스트 데이터 이용

데이터를 양분하여 한 개체 그룹으로부터 판별식을 유도하고, 이 판별식을 사용하여 다른 그룹의 개체를 분류하여 오분류 비율을 추정한다. 표본 자료의 1/2 만 사용하여 판별식을 구하므로 모집단 분류에 적합한 판별식을 얻을 가능성이 낮고 데이터를 많이 수집해야 한다는 단점으로 인하여 이 방법 역시 사용 빈도가 낮다.

(3)Cross-validation 추정법

Lachenbrush(1968)가 제안한 방법으로 가장 널리 사용된다. 첫 번째 개체 하나를 제외하고 판별식을 구하여 그 개체를 분류하고, 첫 번째 개체를 다시 넣고 두 번째 개체를 제외하고 판별식을 구한 후 두 번째 개체를 분류하고..... 이렇게 하여 오분류 비율을 추정한다. 이 방법을 Jackknife 방법이라고도 한다. 모집단이 2 개인 경우 분류표는 다음과 같다.

마취 예제에서 2 개의 판별식에 대해 Cross-validation 방법에 의해 다음 분류표를 얻었다고 하자.

판별식1 ➔	마취 가능	마취 위험	판별식2 ➔	마취 가능	마취 위험
마취 가능	95	10	마취 가능	90	5
마취 위험	5	90	마취 위험	10	90

두 판별식의 오분류 비율은 동일하지만 마취 위험인 환자를 마취 가능 환자로 분류하면 의료 사고 분쟁 소지가 있으므로 이 셀의 오분류 비율이 낮은 판별식 1 이 선호된다. 앞에서 언급 하였듯이 비용 함수 계산이 쉽지 않으므로 현실적으로 동일 비용 함수(equal cost function)나 비례 비용 함수(ratio cost function)가 주로 사용된다.

6.2.3. 예제

미국 Kansas 주립대학 Dr. Michael Finnegan 교수는 야생 칠면조와 사육 칠면조를 구별하기 위하여 수컷 칠면조 82 마리에 대해 9 개 항목에 대한 측정치를 조사하였다. TURKEY.txt/[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p223]

ID: 칠면조 id	HUM: 상완골 길이	RAD: 요골 길이
ULN: 척골) 길이	FEMUR: 대퇴골 길이	TIN: 경골 길이
CAR: carp metacarpus 길이	D3P: 지골까지 길이	COR: 오탁상 길이
SCA: 견갑골 길이	TYPE: 칠면조 종류 야생(WILD), 사육(DOMESTIC)	

```
B790  156  137  151  146  155  814  305  111  137  WILD
B791    .  132  148  138  145  775    .  106  128  WILD
```

우선 판별 분석에 대한 이해를 위하여 **HUM, ULN 두 변수만 측정하였다고** 가정하고 판별 분석을 실시하자. 판별 분석에서는 각 개체의 측정 변수 중 하나라도 결측치가 있으면 그 개체는 분석에서 제외된다.

(1)자료에 대한 산점도

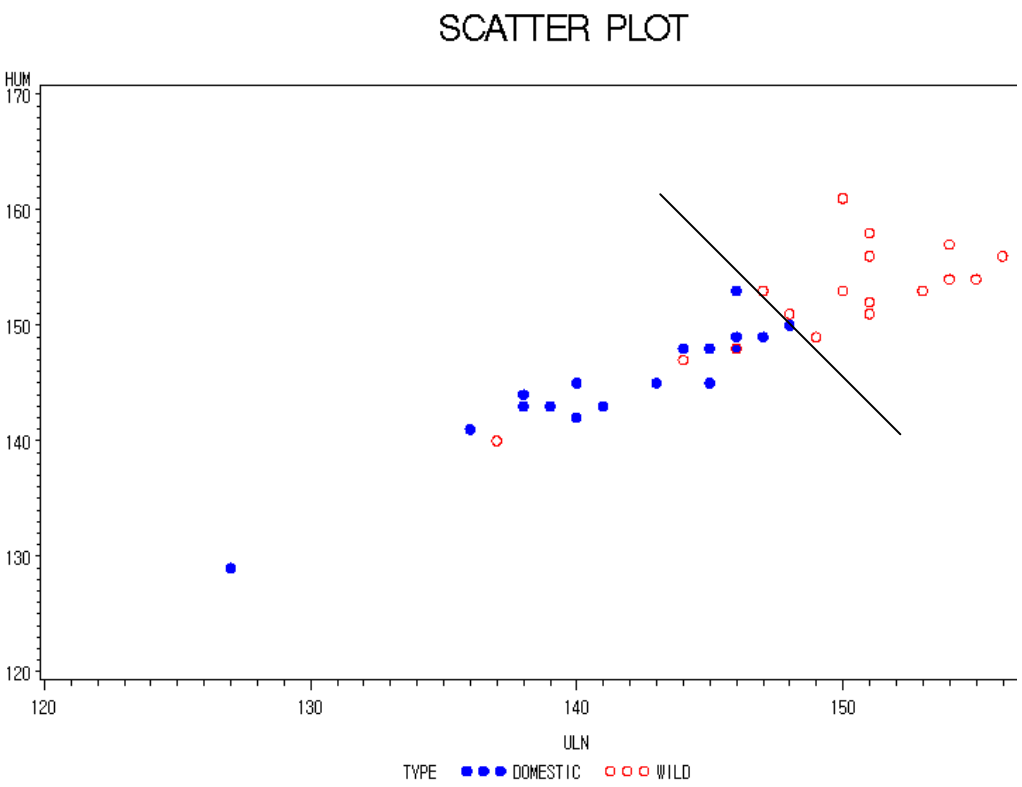
두 변수 HUM(상완골 길이), URN(척골 길이)에 대한 산점도를 칠면조 TYPE 에 따라 산점도를 그리자.

```
DATA TURKEY;
     INFILE 'C:\TEMP\TURKEY.TXT';
     INPUT ID $  HUM RAD ULN FEMUR TIN CAR D3P COR SCA TYPE $;
RUN;

GOPTIONS RESET=ALL;
TITLE 'SCATTER PLOT';
PROC GPLOT DATA=TURKEY;
     SYMBOL1 V=DOT C=BLUE;
     SYMBOL2 V=CIRCLE C=RED;
     PLOT HUM*ULN=TYPE;
RUN;
```

①SYMBOL 문은 점들에 대한 속성을 설정하는 것으로 V 는 Value(점의 속성), C 는 color(점의 색)를 지정한다.

②GOPTIONS 은 그래프 옵션을 지정하는 것이고 RESET=ALL 의 의미는 이전 그래프 옵션 모두를 원래 상태로 돌린다. 가능하면 사용하는 것이 좋다.



만약 이 선을 판별식으로 사용하면 야생 세 마리가 사육으로 오분류 된다.

(2)Fisher 의 판별 분석 프로그램

```
PROC DISCRIM DATA=TURKEY CROSSLIST OUT=OUT1;
    CLASS TYPE;
    VAR HUM ULN;
RUN;
```

①CROSSLIST 옵션은 오분류 비율을 보기 위한 것으로 cross-VALIDATION 방법(가장 많이 사용되는 방법)으로 분류한 것이다. Resubstitution 방법도 함께 출력된다.

②판별 분석 방법은 METHOD=NORMAL 혹은 NPAR 지정할 수 있는데 NORMAL 은 측정 변수들의 다변량 정규 분포 가정 하에서 Fisher 방법을 사용한다는 것을 의미하며 이것이 default 이므로 지정하지 않아도 된다. 다변량 정규 분포 가정이 무너지면 Nonparametric 방법이 사용된다.

③OUT 옵션은 판별 분석 결과를 SAS data 로 저장하기 위한 것이다.

④CLASS 는 개체의 집단 변수를 지정하고 VAR 은 판별 변수를 지정하면 된다.

(3)출력 결과

①일반 사항

The DISCRIM Procedure

Observations	40	DF Total	39
Variables	2	DF Within Classes	38
Classes	2	DF Between Classes	1

▪칠면조 수는 82 마리였지만 HUM, ULN 변수 모두 관측된 칠면조 수는 40 였다. 판별에 사용된 측정 변수는 2 개이고 개체 집단 개수는 2 개임을 나타낸다.

Class Level Information

TYPE	Variable Name	Frequency	Weight	Proportion	Prior Probability
DOMESTIC	DOMESTIC	22	22.0000	0.550000	0.500000
WILD	WILD	18	18.0000	0.450000	0.500000

▪판별식 개체 집단 중 사육에 속한 개체(칠면조)가 22 마리, 야생 칠면조가 18 마리이다.

▪집단 각각의 prior prob. (사전 확률)는 비용 함수와 같이 고려된다. 일반적으로는 균등 확률을 고려하므로 default 는 균등 확률이다. 사전 확률 설정 방법은 나중에 다루기로 한다.

②Fisher 판별식

Linear Discriminant Function

$$\text{Constant} = -.5\,\bar{X}_j'\,COV^{-1}\,\bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1}\,\bar{X}_j$$

상수 부분 ↑ ↑ 계수 부분

Linear Discriminant Function for TYPE

Variable	DOMESTIC	WILD
Constant	-484.19852	-533.56136
HUM	3.86887	3.95324
ULN	2.84207	3.09356

▪상수(constant) $k = (1/2)(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 + \underline{\mu}_2)$ / 계수(coefficient) 부분 $\underline{b}' = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$

▪판별 함수 (classification function): $\underline{b}'\underline{x}_0 - k$ 의 부호가 양수이면 π_1 으로 분류하고, 그렇지 않으면 π_2 으로 분류한다. 즉 각 개체에 대해 변수들에 대해 $\underline{b}'\underline{x}_0$ 을 구한 후 가장 큰 값을 갖는 집단으로 분류하면 된다.

③판별 분석 분류 결과

다음은 Resubstitution 방법에 의한 분류표이다.

Classification Summary for Calibration Data: WORK.TURKEY
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$D^2_{(X)} = (X - \bar{X}_{(\cdot)})' COV^{-1} (X - \bar{X}_{(\cdot)})$

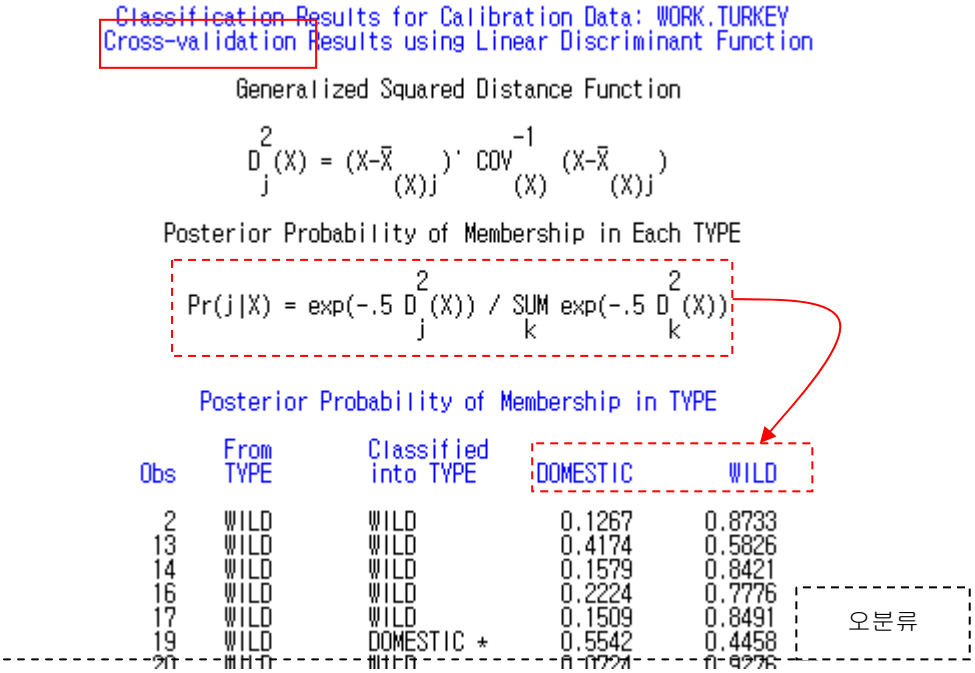
Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	16 72.73	6 27.27	22 100.00
WILD	3 16.67	15 83.33	18 100.00
Total	19 47.50	21 52.50	40 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.2727	0.1667	0.2197
Priors	0.5000	0.5000	

이미 언급하였듯이 Re-substitution 오 분류 비율이 under-estimate 되므로 사용하지 않는다.



▪19 번째 개체는 원래 야생 칠면조인데 Fisher 판별식에 의해 사육 칠면조로 분류되었으므로 오분류에 해당된다. From type 은 원래 집단이고 Classified into 는 판별식에 의해 분류된 집단을 의미한다.

▪Posterior Prob. (사후 확률)은 개체가 각 집단(그룹)에 속할 확률이므로 사후 확률이 큰 집단으로 분류된다. 2 번째(Obs=2) 개체는 사육 칠면조일 사후 확률이 0.13, 야생 칠면조일 사후 확률일 0.87 이므로 야생 칠면조에 분류된다. 두 사후 확률의 합은 당연히 1 이다.

④오분류 표

Cross-validation Summary using Linear Discriminant Function
Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
	1 50.00	1 50.00	2 100.00
DOMESTIC	16 72.73	6 27.27	22 100.00
WILD	3 16.67	15 83.33	18 100.00
Total	20 47.62	22 52.38	42 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.2727	0.1667	0.2197
Priors	0.5000	0.5000	

▪원은 오분류, 사각형은 정분류를 의미한다.

▪Fisher 판별식, cross-validation 방법에 의한 오분류 결과가 정리된다.

▪사육 ▶ 야생으로 오분류된 개체 수가 6 개이므로 오분류 비율은 6/22=0.2727 이다.
야생 ▶ 사육으로 오분류된 개체 수가 3 개이므로 오분류 비율은 3/18=0.1667 이다.

▪총 오분류 비율은 (0.2727+0.1667)/2=0.2197 이다.

▪오분류 비용 함수가 균등하다면 총 오분류 비율이 낮은 판별 분석 방법을 선택하면 된다.

⑤분류 결과 보기

판별 분석 결과를 SAS data OUT1 에 저장하였으므로 이를 출력해 보자. (OUT=OUT1 옵션)

```
PROC PRINT DATA=OUT1;
RUN;
```

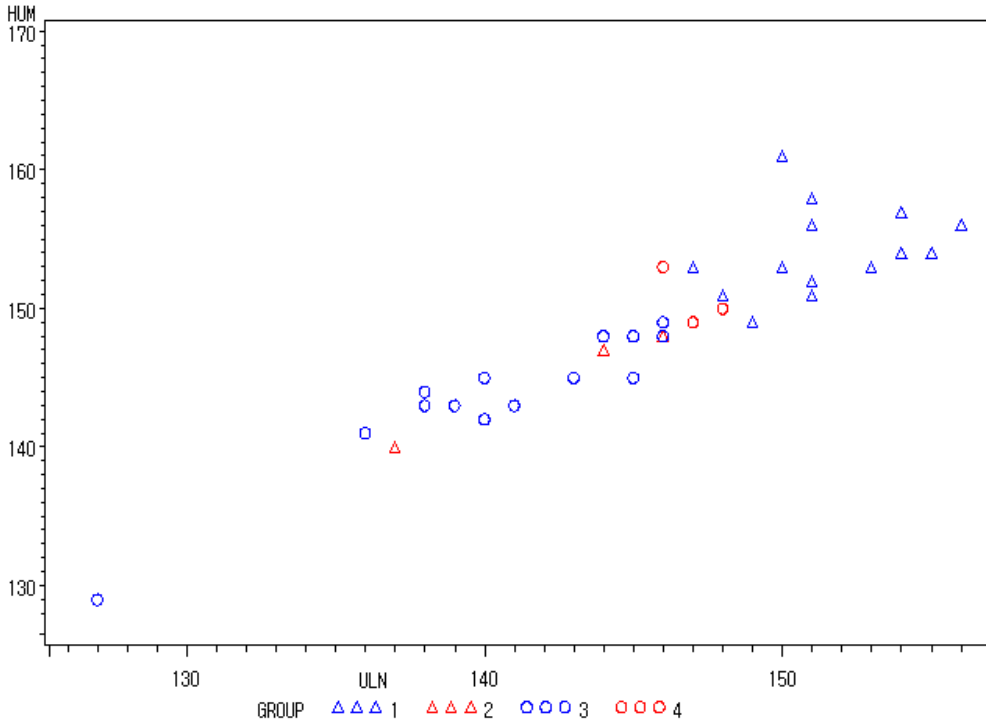
Obs	ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	DOMESTIC	WILD	_INTO_
1	K766	.	.	.	.	.	.	.	.	.	WILD	.	.	
2	N399	153	138	153	139	162	810	307	.	.	WILD	0.11680	0.88320	WILD
3	NEX1	.	.	.	.	.	.	.	.	.	WILD	.	.	

분류 결과는 _INTO_라는 자동 생성 변수에 저장되어 있다. 사용되지 않은 개체 분류는 공백이므로 산점도를 그리기 위해 제외하고 분류 결과의 산점도를 그려보자.

```
DATA OUT2;
    SET OUT1;
    IF (TYPE="WILD") AND (_INTO_="WILD") THEN GROUP=1;
    IF (TYPE="WILD") AND (_INTO_="DOMESTIC") THEN GROUP=2;
    IF (TYPE="DOMESTIC") AND (_INTO_="DOMESTIC") THEN GROUP=3;
    IF (TYPE="DOMESTIC") AND (_INTO_="WILD") THEN GROUP=4;
    IF (GROUP^=.);
RUN;

GOPTIONS RESET=ALL;
TITLE 'SCATTER PLOT';
PROC GPLOT DATA=OUT2;
    SYMBOL1 V=TRIANGLE C=BLUE;
    SYMBOL2 V=TRIANGLE C=RED;
    SYMBOL3 V=CIRCLE C=BLUE;
    SYMBOL4 V=CIRCLE C=RED;
    PLOT HUM*ULN=GROUP;
RUN;
```

판별되지 않은 개체를 제외하기 위하여 IF (GROUP^=.)을 사용하였다. 정분류는 파랑색, 오분류는 빨간 색으로 표시하였다.



빨간 색이 오분류이다. 빨간 삼각형은 사육 칠면조인데 야생 칠면조로 분류된 것이고 (3 마리) 빨간 원은 야생이 사육으로 분류된 오분류이다. 6 개인데 겹쳐져 3 개처럼 보인다.(실제로 프로그램을 실행해 보면 차이를 알 수 있을 것이다.)

(4)새로운 개체 분류하기

새로운 칠면조에 2 마리 왔는데 야생 칠면조인지 사육 칠면조인지 알 수 없어 판별하고자 한다. 두 마리의 (HUM, ULN)을 측정하였더니 다음과 같았다.

(HUM, ULN) = (145, 150) , (HUM, ULN) = (150, 145)

임시 SAS 데이터를 만든다. OUTPUT 은 현재 값을 데이터에 저장하라는 의미이므로 NEW 에는 2 개 관측치가 있고 HUM, ULN 변수 제외하고는 모두 결측치이다. 개체 그룹을 나타내는 TYPE 도 결측치이므로 판별식 구하는데 사용되지 않지만 사후 확률은 계산되므로 판별은 된다.

```
DATA NEW;
     HUM=145; ULN=150; OUTPUT;
     HUM=150; ULN=145; OUTPUT;
RUN;

DATA FIN;
     SET TURKEY NEW;
RUN;

PROC DISCRIM DATA=FIN CROSSLIST;
     CLASS TYPE;
     VAR HUM ULN;
RUN;
```

(HUM, ULN) = (145, 150)인 칠면조는 야생이 확률이 0.64 로 크므로 야생으로 분류되고 (HUM, ULN) = (150, 145) 칠면조는 사육일 확률이 0.56 이므로 사육으로 분류된다. (사후 확률 0.5 기준)

64	DOMESTIC	WILD	*	0.3576	0.6424
83		WILD	*	0.3558	0.6442
84		DOMESTIC	*	0.5602	0.4398

새로운 개체 2 개는 판별식 구하는데 이용되지 않았으므로 오분류 표는 이전과 동일하다. 다른 것이 있다면 새로운 행이 하나 더 생겨 새로운 개체 판별 결과도 정리된다.

Cross-validation Summary using Linear Discriminant Function
Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
비어 있음	1 50.00	1 50.00	2 100.00
DOMESTIC	16 72.73	6 27.27	22 100.00
WILD	3 16.67	15 83.33	18 100.00
Total	20 47.62	22 52.38	42 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.2727	0.1667	0.2197
Priors	0.5000	0.5000	

6.3. 알아 두기

6.3.1. 오분류 결과 저장하기

판별식으로부터 오분류를 계산하는 방법은 3 가지가 있다고 앞에서 설명하였다. Resubstitution, Test Data 이용, Cross-Validation. SAS 의 출력 결과 중 Posterior Prob.는 Cross-validation 방법에 의한 분류에 사용되는 사후 확률이다. 이 결과를 SAS data 로 저장하려면 OUTCROSS=이름 옵션을 사용해야 한다. 만약 OUT=이름(예:OUT1) 옵션을 사용하면 Resubstitution 방법 분류를 위한 사후 확률 값과 결과가 저장된다. Test data 는 OUTTEST=이름 옵션을 사용하면 된다.

```
PROC DISCRIM DATA=TURKEY CROSSLIST OUTCROSS=OUT1;
    CLASS TYPE;
    VAR HUM ULN;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

Obs	ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	DOMESTIC	WILD	_INTO_
1	K766	.	.	.	.	.	.	.	.	.	WILD	.	.	
2	N399	153	138	153	139	162	810	307	.	.	WILD	0.12673	0.87327	WILD
3	NFX1	.	.	.	.	.	.	.	.	.	WILD	.	.	
16	B795	151	134	151	144	.	789	292	116	126	WILD	0.22235	0.77765	WILD
17	B819	158	135	151	146	152	790	289	111	125	WILD	0.15087	0.84913	WILD
18	B081	.	135	149	.	149	789	.	111	123	WILD	.	.	
19	B085	148	129	146	139	147	767	287	106	123	WILD	0.55423	0.44577	DOMESTIC
20	B089	157	140	154	140	159	818	301	116	136	WILD	0.07237	0.92763	WILD

6.3.2. 오분류 비용 함수와 사전 확률

모집단이 2 개인 경우 오분류 비용 함수를 다음과 같이 정의하자.

(1) $C_{2|1}$: 1 집단을 2 집단으로 분류했을 때 발생하는 비용

(2) $C_{1|2}$: 2 집단을 1 집단으로 분류했을 때 발생하는 비용

또한 각 집단의 사전 비율을 알고 있다면 이를 판별식 유도에 사용할 수 있을 것이다. 모집단이 2 개인 데이터에서 1 집단의 사전 비율을 p_1, 2 집단의 사전 비율을 p_2 라 하자.

$$d_i^* = 1/2(\underline{x}_0 - \underline{\mu}_i)' \Sigma^{-1} (\underline{x}_0 - \underline{\mu}_i) - \ln(p_i^*), \quad i = 1,2$$

$$p_1^* = \frac{p_1 C(2|1)}{p_1 C(2|1) + p_2 C(1|2)}, \quad p_2^* = \frac{p_2 C(1|2)}{p_1 C(2|1) + p_2 C(1|2)}$$

만약 $d_1^* > d_2^*$ 이면 1 집단(π_1)으로 분류되고 $d_1^* < d_2^*$ 이면 2 집단(π_2)으로 분류된다. 칠면조 예에서 Wild 값과 Domestic 값 중 큰 값을 갖는 집단에 분류하는 것과 같다. 단 여기서는 비용과 사전 확률을 고려하는 차이만 있다.

사전 확률만 고려할 수도(서로 다른 p1 과 p2 사용하고 비용 함수는 다르게 $C(1|2) = C(2|1)$), 오분류 비용 함수만 다르게 고려할 수도($p_1 = p_2$) 있다. 그러나 비용 함수를 고려하는 것은 매우 어려우므로 비용 함수는 동일하다는 가정을 하게 된다.

사전 확률을 옵션을 사용하는 경우는 판별에 사용되는 데이터의 집단 구성 비율이 모집단의 비율과 현저히 다르고 모집단 비율을 알고 있을 때이다. 옵션은 다음과 같다.

(1)PRIORS EQUAL

모집단의 집단 크기가 동일할 때 (default)

(2)PRIORS PROPORTIONAL

표본 데이터 집단 구성 비를 그대로 사용할 때

(3)PRIORS 'WILD'=0.4 'DOMESTIC'=0.6

사육 칠면조 0.6, 야생이 0.4 인 사전 정보가 있을 때

```
PROC DISCRIM DATA=TURKEY CROSSLIST OUTCROSS=OUT1;
    CLASS TYPE;
    VAR HUM ULN;
    PRIORS 'WILD'=0.4 'DOMESTIC'=0.6;
RUN;
```

6.2.3 절의 오분류 결과와 비교해 보면 사육=>야생 오분류는 6->2 개로 4 개 줄고, 야생=>사육 오분류는 3 개에서 4 개로 한 개 늘어 전체 오분류는 줄었다. 이처럼 집단 구성 비율에 대한 사전 정보가 있으면 오분류를 줄일 수 있다. 물론 사전 정보가 잘못되면 오분류가 늘어날 수도 있지만.....

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	20 90.91	2 9.09	22 100.00
WILD	4 22.22	14 77.78	18 100.00
Total	24 60.00	16 40.00	40 100.00
Priors	0.6	0.4	

6.3.3. 등분산 가정 검정하기

SAS 가 판별 분석 진행 과정에서 각 변수들의 집단간 등분산 가정을 검정하여 채택되면 통합 분산-공분산을 사용하고 그렇지 않으면 자동으로 집단 내(within) 분산-공분산을 사용하므로 판별 분석에서 등분산 가정을 걱정할 필요는 없다. 만약 분석자가 임의로 통합 분산-공분산을 사용하고 싶으면 POOL=YES 옵션 사용하면 된다. 다음은 등분산 가정을 검정하는 방법이다.

```
PROC DISCRIM DATA=TURKEY CROSSLIST POOL=TEST;
    CLASS TYPE;
    VAR HUM ULN;
RUN;
```

Under the null hypothesis: $-2\ RHO\ \ln\left[\frac{N^{PN/2}}{\prod N(i)^{PN(i)/2}}\right]$

is distributed approximately as Chi-Square(DF).

Chi-Square	DF	Pr > ChiSq
7.020815	3	0.0712

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252

등분산 가정을 검정 결과에 집단간 분산의 차이가 있으므로(유의수준 0.1) 집단 내 공분산이 판별식 유도에 사용되었다.

6.4. 단계적 판별 분석

판별 분석에서 많은 변수(항목)들이 측정되었을 경우 아마 여러분은 의문이 생길 것이다. (1)모든 변수가 판별에 필요한가? (2)어떤 변수가 가장 판별을 잘하는 변수인가? 결론을 말하자면 판별에 적절한 변수만 사용하는 것이 좋으며 판별 능력은 분산 분석의 개념을 이용하여 판단한다. 판별 변수 선택 방법으로는 회귀 분석과 유사하게 Forward 방법, Backward 방법, Stepwise 방법이 있다.

6.4.1. Forward 방법

다음 절차에 의해 개체를 판별하는데 가장 유의한 변수 순으로 유의한 변수가 존재하지 않을 때까지 하나씩 넣어 가는 방법이다.

(1)개체 집단을 설명 변수(요인)로 하고 각 측정 변수를 종속 변수(반응 변수)로 하여 분산 분석(ANOVA)을 실시한다. F-값이 가장 큰 변수를 제일 먼저 선택한다. 이유는 집단의 평균 차이가 가장 크다는 것은 그 변수에 의해 집단 분류가 가장 잘된다는 것이다.

(2)두 번째 변수 선택은? 첫 번째 선택된 변수를 공변량(covariate)으로 하여 공분산 분석 (ANCOVA) 시행하여 그룹의 SS3 F-값이 가장 큰 변수를 선택한다. 공변량은 종속 변수에서 그 변수의 효과를 제외할 때 사용되므로 첫 번째 선택된 변수의 판별 효과를 제외하는 것을 의미한다.

🕮 **공분산 분석:** 수학 강의에 대한 새로운 교육방법이 제안되었다. 기존의 교육 방법보다 나은지 알아보기 위하여 각 20 명씩 2 개의 그룹을 만들어 하나의 그룹에는 새로운 교육 방법, 다른 그룹에는 기존의 교육 방법을 적용해 보자. 그룹 학생들간에는 차이가 있을 것을 예상하여 교육 전 수학 시험을 보았다. 일전 기간 교육 후 수학 능력 시험을 봐 그 성적의 차이가 있는지 분석하였다. 교육 후 점수(Y)가 그룹(새 교육/기존 교육)간 차이가 있는지 알아보려면 분산분석(ANOVA). 그러나 교육 전 이들의 수학 능력이 고려되지 않았다. 모두 수학에 대한 능력이 같지는 않을 것이다. 이 효과를 제외해 주자. 이 역할을 하는 것이 교육 전 수학 점수이고 이를 공변량이라 한다. 이에 적합한 분석이 공변량 분석이다. 여전히 주요 관심은 교육 효과이고 공변량에는 관심이 없다.

세 번째 변수 선택은? 첫 번째, 두 번째 선택된 변수들을 공변량(covariate)으로 하여 공분산 분석 (ANCOVA) 시행하여 그룹의 SS3 F-값이 가장 큰 변수를 선택한다. 이렇게 변수 선택을 반복한다. 만약 F-값이 가장 큰 것이 유의하지 않으면 (SS3 의 p-값이 유의수준보다 크면) 변수 선택을 멈춘다. 일반적으로 유의수준은 0.25 와 0.5 사이로 한다. SAS 에서는 SLE(Significant Level for Entry) 옵션을 설정할 수 있다.

6.4.2. Backward 방법

다음 절차에 의해 개체를 판별하는데 유의한 변수를 선택하는 방법으로 일단 모든 변수를 다 고려한 후 유의하지 않은 순서대로 변수를 제거해 나가는 방법이다.

(1)하나의 변수를 반응 변수, 다른 변수들은 공변량, 그리고 그룹을 요인(설명 변수)으로 하여 공분산 분석을 실시하여 집단의 (Type III, Partial SS) F-값이 가장 낮은 변수를 제거한다.

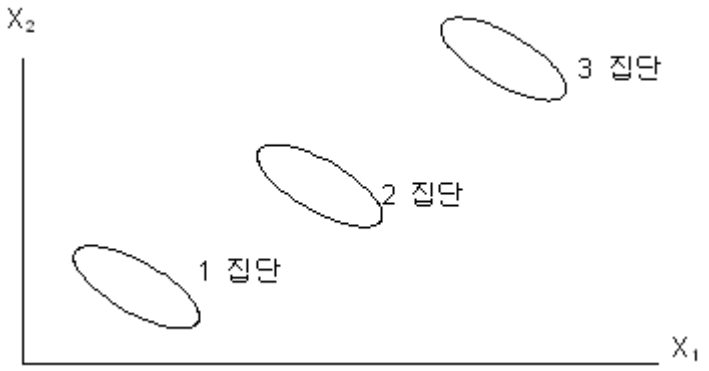
(2)같은 방법으로 변수를 하나씩 제거해 간다. 집단의 SS3 의 F-값이 모두 유의하면 (p-값이 유의수준보다 작으면) 제거를 멈춘다. 일반적으로 유의수준은 0.15 로 한다. SAS 에서는 SLS(Significant Level for Stay) 옵션을 설정할 수 있다.

6.4.3. Stepwise 방법

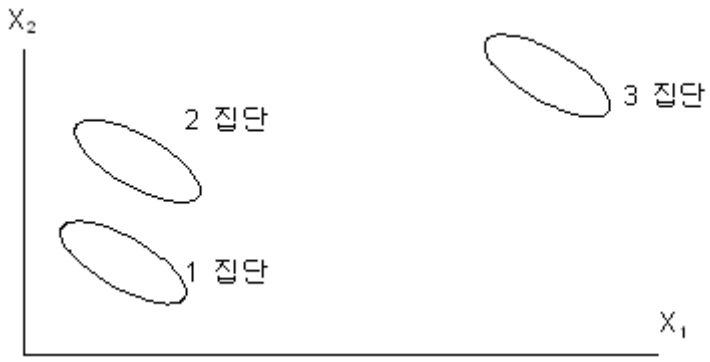
Forward 방법과 매우 유사하다. 일단 선택된 변수들도 다른 변수가 들어간 상태에서 유의성 검정을 하여 새로운 변수보다 덜 유의하면 제거된다. 즉 처음에는 가장 유의하였지만 여러 변수들이 선택된 상황에서는 유의 정도가 떨어질 수 있다. 변수 선택을 위하여 변수를 넣었다 뺐다 하는 반복이 많아 계산이 복잡하고 번거롭지만 컴퓨터 하드웨어, 소프트웨어 발달로 인하여 현재는 가장 많이 사용되는 방법이다. SAS 에서는 SLS, SLE 옵션을 설정할 수 있다.

6.4.4. 변수 선택 시 주의점

변수의 수가 15 개 이상인 경우 Backward 방법, 15 개 미만이 경우는 Stepwise 방법을 사용하는 것을 권한다. 판별 분석에서 변수 선택은 다음과 같은 문제점을 지니고 있다. 판별에 유의한 변수를 찾는 경우 F-값만 가지고 선택하므로 그 변수가 얼마나 잘 판별하는지를 고려되지 않는다. 두 변수 (X_1, X_2) 대해 집단간 산점도가 아래 그림과 같다면 변수 X1 이 집단을 가장 잘 분류하므로 먼저 선택되고 X2 가 그 다음으로 선택된다. 모두 유의하다. 별 문제가 없어 보인다.



만약 산점도가 위와 같다면 실제로는 X_2 가 더 나은 판별을 하지만 변수 선택 방법으로 선택하면 X_1 이 선택될 것이다. 그러므로 변수 선택 방법에 의해 선택된 변수가 판별을 잘한다는 보장은 없다. 그럴지라도 변수 선택 방법은 하나의 좋은 기준을 제시한다.



판별 분석의 주요 목적은 개체들을 오분류 없이 분류하는 것이므로 다양한 변수 선택 방법을 사용해 보고 각각에 대해 (1)측정 변수 모두 사용하여 오분류 비율을 구하고 (2)변수 선택 후 오분류 비율을 계산하여 오분류가 가장 적은 판별 분석 방법과 측정 변수 군을 이용하는 것이 올바른 접근 방법이다.

6.4.5. 변수 판별의 필요성

오분류 비율이 동일하거나 차이가 없다면 판별 변수의 수가 적은 판별 방법이 더 효율적이다. 이유는 측정 오류 발생 가능성이 적으며 새로운 개체 판별을 위해 측정해야 하는 변수 수가 적으므로 경제적이다. 판별 변수 개수에 따른 오분류 결과를 비교하기 위하여 모든 측정 변수가 다 관측되어 있는 칠면조만 판별 분석에 사용되었다. (**TURKEY0.TXT**)

(1)측정 변수 모두 사용하기

```
DATA TURKEYO;
    INFILE 'C:\TEMP\TURKEYO.TXT';
    INPUT ID $ HUM RAD ULN FEMUR TIN CAR D3P COR SCA TYPE $;
RUN;


PROC DISCRIM DATA=TURKEYO CROSSLIST OUTCROSS=OUT1;
    CLASS TYPE;
    VAR HUM RAD ULN FEMUR TIN CAR D3P COR SCA;
RUN;
```

개체 분류 결과는 CROSSLIST 에 의해서도 출력이 되며 그 결과는 저장하고 싶다면 OUTCROSS 옵션을 사용하면 된다. 측정된 8 개 변수 모두를 사용하였고 cross-validation 결과를(분류, 사후 확률) OUT1 에 저장하였다.

The DISCRIM Procedure

Observations	32	DF Total	31
Variables	9	DF Within Classes	30
Classes	2	DF Between Classes	1

Class Level Information

TYPE	Variable Name	Frequency	Weight	Proportion	Prior Probability
DOMESTIC	DOMESTIC	19	19.0000	0.593750	0.500000
WILD	WILD	13	13.0000	0.406250	0.500000

9 개 변수 모두 측정된 칠면조 수가 32 개이고 그 중 19 마리가 사육, 13 마리가 야생이다. PRIOR 옵션을 사용하지 않았으므로 집단 비율은 equal (0.5) 사용하였다.

Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	17 89.47	2 10.53	19 100.00
WILD	1 7.69	12 92.31	13 100.00
Total	18 56.25	14 43.75	32 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.1053	0.0769	0.0911
Priors	0.5000	0.5000	

(HUM, ULN) 두 변수를 사용하였을 의 오분류 표는 다음과 같다. 6.2.3 절 결과와 다른 이유는 측정 변수 9 개 모두 측정된 칠면조만 판별에 이용하였기 때문이다.

Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	17 89.47	2 10.53	19 100.00
WILD	2 15.38	11 84.62	13 100.00
Total	19 59.38	13 40.63	32 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.1053	0.1538	0.1296
Priors	0.5000	0.5000	

(2)변수 선택하여 분류하기

판별 변수 선택 방법에는 Forward, Backward, Stepwise 방법이 있는데 그 중 가장 많이 이용되는 방법이 Stepwise 이므로 이 방법을 이용하여 개체 분류에 유의한 변수를 선택한 후 Fisher 판별식으로 개체를 분류하여 보자.

```
PROC STEPDISC DATA=TURKEY0 METHOD=STEPWISE SLE=0.25 SLS=0.15;
    CLASS TYPE;
    VAR HUM ULN FEMUR TIN CAR D3P COR SCA;
RUN;
```

변수를 포함시킬 때 유의수준은 0.25-0.4, 제거할 때 유의수준은 0.15 가 적절하다고 한다. 여기서는 포함시킬 때 유의 수준을 0.25, 제거할 때 유의 수준을 0.15 로 하여 Stepwise 방법으로 변수를 선택한 후 판별 분석을 실시한 프로그램이다. SAS 에서 판별 변수 선택은 **STEPDISC** procedure 를 사용하면 된다.

er In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	TIN		0.7127	74.42	<.0001	0.28730847	<.0001	0.71269153	<.0001
2	COR		0.1648	5.72	0.0234	0.23994800	<.0001	0.76005200	<.0001
3	D3P		0.1371	4.45	0.0440	0.20705038	<.0001	0.79294962	<.0001
4	ULN		0.1009	3.03	0.0930	0.18614907	<.0001	0.81385093	<.0001

출력 결과 마지막에 나오는 변수 선택 요약을 보면 된다. 판별 선택 변수는 TIN, COR, D3P, ULN 이다. 이를 이용하여 Fisher 판별 분석을 실시하여 보자.

```
PROC DISCRIM DATA=TURKEY CROSSLIST OUTCROSS=OUT2;
    CLASS TYPE;
    VAR TIN COR ULN D3P;
RUN;
```

Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	17 89.47	2 10.53	19 100.00
WILD	1 7.69	12 92.31	13 100.00
Total	18 56.25	14 43.75	32 100.00
Priors	0.5	0.5	

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.1053	0.0769	0.0911
Priors	0.5000	0.5000	

4 개 변수만 사용하여 판별하였을 경우 사육=>야생, 야생=>사육 오분류 비율 동일하다. 그러므로 변수 선택하여 판별 변수 수를 줄이는 것이 효율적이다. 측정 변수가 많을수록 측정 비용이 많이 들고 측정 오류 발생 가능성이 높아지기 대문이다.

(3)오분류 결과 보기

판별 변수 4 개를 사용하나 9 개를 사용하나 오분류 비율도 동일하고 오분류 되는 개체도 동일하다. 어떤 칠면조가 오분류 되었는지 알아보기 위하여 오분류 칠면조만 살펴보기로 하자. 선택된 4 개 변수 사용하였다. 오분류 개체(칠면조)만 출력하기 위하여 OUT2 에 TYPE 과 _INTO_가 다른 개체를 저장하였다. 그리고 그 개체들이 왜 오분류 되었는지 알아보기 위하여 각 집단의 측정 변수 평균을 구해 보았다. 물론 오분류 개체는 제외하는 것이 좋다.

```
DATA OUT2;
    SET OUT1;
    IF (TYPE^=_INTO_);
RUN;
PROC PRINT DATA=OUT2;
    VAR ID TIN COR D3P ULN TYPE _INTO_;
RUN;
```

Obs	ID	TIN	COR	D3P	ULN	TYPE	_INTO_
1	B710	151	102	305	147	WILD	DOMESTIC
2	L684	146	107	310	146	DOMESTIC	WILD
3	L750	147	104	300	147	DOMESTIC	WILD

칠면조 B710 의 경우 COR 에 의해 L684 과 L750 은 4 개 측정 변수에 의해 오분류 되었다. 그러므로 다른 측정 변수는 WILD 에 가까운데 COR 만 큰 칠면조는 사육으로 분류될 수 있으므로 새로운 개체 판결에 주의해야 한다

```
DATA OUT3;
    SET OUT1;
    IF (TYPE=_INTO_);
RUN;
PROC TABULATE DATA=OUT3;
    CLASS TYPE;
    VAR TIN COR D3P ULN;
    TABLE TYPE, (TIN COR D3P ULN)*MEAN;
RUN;
```

	TIN	COR	D3P	ULN
	Mean	Mean	Mean	Mean
TYPE				
DOMESTIC	137.71	99.53	295.88	143.47
WILD	152.87	110.13	297.87	151.27

6.4.6. 새로운 개체 분류하기

판별에 사용되는 변수의 개수 차이가 있는 것을 제외하고는 새로운 개체 판별 방법은 동일하다. 새로운 칠면조의 (TIN, COR, D3P, ULN) = (140, 105, 300, 145)였다고 하자.

```
DATA NEW;
    TIN=140;COR=105; D3P=300;ULN=145;OUTPUT;
RUN;

DATA ALL;
    SET TURKEY NEW;
RUN;

PROC DISCRIM DATA=ALL CROSSLIST;
    CLASS TYPE;
    VAR TIN COR D3P ULN;
RUN;
```

새로운 칠면조는 사육(domestic) 사후 확률이 0.957 이므로 사육으로 판별된다.

```
63    DOMESTIC    DOMESTIC        1.0000      0.0000
64    DOMESTIC    DOMESTIC        0.9857      0.0143
83                DOMESTIC *      0.9568      0.0432
```

6.5. 정준 판별 분석

Fisher 에 의해 제안된 방법으로 Fisher's between-within method 라고 불리는 방법이다. 판별 변수들의 유용한 정보를 모두를 포함한 정준 (Canonical) 변수를 이용하여 판별 분석을 실시한다. 판별 변수들의 수가(p) 너무 많아 판별 결과에 대한 해석이 곤란한 경우 p -차원 공간에서의 개체들의 집단 평균들을 저 차원 공간으로 변환시켜 처리하는 판별 분석 방법이다. 개체 분류가 목적이 아니라 개체 분류 해석을 위해 저 차원(BOX-PLOT 이나 산점도)으로 표현하는데 있으므로 엄밀히 말하면 판별 분석은 아니다. 새로운 변수(정준 변수)에 대한 해석이 가능하든 아니든 집단들 사이의 실제 거리를 저 차원으로 축소하여 시각화 할 수 있다는 장점이 있다. 차원을 줄인다는 의미에서 보면 주성분 분석과 유사해 보이지만 계산 방법은 전혀 다르다.

6.5.1. 제일 정준 함수

먼저 다음과 같은 가정을 하자.

$\pi_i \sim N_p(\underline{\mu}_i, \Sigma)$ 서 표본을 각각 n_i 뽑았다고 하자. $i = 1,2,...,m$ (m 개 모집단)각 모집단은 차수 p 인 다변량 정규분포를 따르면 동일한 분산-공분산 행렬을 갖는다.

▪Between sample mean: $B = \sum_{i=1}^{m} n_i (\underline{\hat{\mu}}_i - \underline{\hat{\mu}}_o)(\underline{\hat{\mu}}_i - \underline{\hat{\mu}}_o)'$, $\underline{\hat{\mu}}_o = \frac{1}{n_o} \sum_{i=1}^{m} n_i \underline{\mu}_i$, $n_o = \sum_{i=1}^{m} n_i$

▪Within sample mean: $W = \sum_{i=1}^{m} \sum_{r=1}^{n_i} (\underline{x}_{ri} - \underline{\hat{\mu}}_i)(\underline{x}_{ri} - \underline{\hat{\mu}}_i)'$

$\max_{\underline{b} \neq 0} \frac{\underline{b}'B\underline{b}}{\underline{b}'(B+W)\underline{b}}$ 을 만족하는 선형 계수 $\underline{b}'$ 는 $(B+W)^{-1}B$ 의 가장 큰 고유치로부터 얻은 고유벡터이다. 이를 $\underline{b}_1$ 이라 하자. $\underline{y}_1 = \underline{b}_1' \underline{x}$ 은 주성분과 동일하다. 각 개체의 집단 평균과의 거리는 $d_i = |\underline{b}_1' \underline{x} - \underline{b}_1' \underline{\hat{\mu}_i}|$, $i = 1,2,...,m$ 이다.

6.5.2. 제이 정준 함수

$\underline{b}_2$ 는 $(B+W)^{-1}B$ 의 두 번째 큰 고유치로부터 구한 고유 벡터이다. $(B+W)^{-1}B$ 로부터 구해진 제 2 주성분과 동일하다. 주 성분 변수가 2 개일 경우 각 개체들과 집단 평균과의 거리는 다음과 같이 계산된다. $d_i = (\underline{b}_1' \underline{x} - \underline{b}_1' \underline{\hat{\mu}_i})^2 + (\underline{b}_2' \underline{x} - \underline{b}_2' \underline{\hat{\mu}_i})^2$, $i = 1,2,...,m$

유사한 방법으로 제삼, 제사 정준 함수를 구할 수 있으나 정준 판별 분석은 다변량 데이터의 개체를 저 차원(산점도)에 나타내는 것이 주목적이다. 따라서 일반적으로 제이 정준 함수 까지만 이용하게 된다.

6.5.3. 차수 결정

주성분 분석과 마찬가지로 누적 설명력(이 80% 이상이 되거나 SCREE plot 에 의해 갑자기 설명력이 뚝 떨어지는 곳까지 선택하면 된다. 일반적으로 p 차원을 저 차원으로 줄이는 것이 목적이므로 2 차까지만 한다.

6.5.4. 예제

정준 판별 분석은 PROC CANDISC 를 이용하면 된다. 8 개 변수 모두를 사용하여 정준 상관 분석을 실시하여 보자. (다시 강조하지만 정준 상관 분석은 저 차원 공간에 개체를 나타내는 것이다.) NCAN 옵션은 정준 변수 개수를 지정하는 것이다. NCAN=2 에 의해 정준 변수 개수를 2 개로 지정하였으므로 OUT 의 CANSCORE SAS 데이터는 CAN1 과 CAN2 가 저장된다.

```
PROC CANDISC OUT=CANSCORE DATA=TURKEY NCAN=2;
    CLASS TYPE;
    VAR HUM RAD ULN FEMUR TIN CAR D3P COR SCA;
RUN;
```

(1) 고유치 개수

The CANDISC Procedure

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.879842	0.850856	0.039930	0.774122

$$\max_{\underline{b}\neq 0} \frac{\underline{b}' B \underline{b}}{\underline{b}' (B+W) \underline{b}}$$

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	3.4272		1.0000	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.22587846	8.76	9	23	<.0001

옵션은 NCAN=2 로 했지만 고유치 하나로 누적 변동 100% 이상이므로 하나만 출력되었다. 고유치의 개수가 적절한지 유의성을 검정한다. 고유치가 하나 밖에(누적 설명이 100%이다) 없으므로 한 개만 유의성 검정한다. 위의 귀무가설은 현재 정준 변수와 그 뒤의 정준 변수 간의 상관 관계에 대한 유의성 검정한다. 만약 현재 정준 변수와 나머지 뒤의 정준 변수와 상관 관계가 유의하면 현재 정준 변수만으로 개체 판별이 가능함을 의미한다.

만약 고유치 개수가 2 개이면 다음과 같이 출력된다.

	Eigenvalue	Difference	Proportion	Cumulative
1	32.1919	31.9065	0.9912	0.9912
2	0.2854		0.0088	1.0000

Test of H0: The canonical correlations in the current row and all that fol

	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.02343863	199.15	8	288	<.0001
2	0.77797337	13.79	3	145	<.0001

정준 변수 개수 2 개에 대한 유의 확률이 <0.001 이므로 정준 변수가 2 개는 필요하다. 각 열의 귀무 가설은 “정준 변수의 개수가 열의 숫자이면 충분하다” 이고, 대립 가설은 “정준 변수의 개수가 열의 숫자보다 커야 한다” 이다.

(2)정준 변수 계수

다음은 정준 변수를 구하는 계수가 출력된 것이다. 각 개체에 대한 제일 정준과 제이 정준 등을 구하는 계수이다(6.5.1 절과 6.5.2 절 참고)

Raw Canonical Coefficients

Variable	Can1	Can2
HUM	-.0288709867	-.1172493125
RAD	-.0220335241	0.1906798839
ULN	-.1706330508	-.0162182652
FEMUR	-.0571149945	0.2366560814
TIN	0.2513770115	-.1988078595
CAR	-.0089556372	-.0205136437
D3P	-.0284160262	0.0021504708
COR	0.2263163950	0.1072236911
SCA	-.0033058634	-.0234610998

개체 13 의 Can1 변수, Can2 변수를 구하는 식을 보자.

$$Can1 = -0.029*153 - 0.022*140 + ... - 0.0033*128 = 0.2016$$
$$Can2 = -0.1172*153 + 0.1906*140 + ... - 0.0234*128 = 0.3969$$

(3)각 집단 정준 변수 평균

사육, 야생 칠면조 집단의 제일 정준 변수의 평균을 구하면 다음과 같다.

Class Means on Canonical Variables

TYPE	Can1
DOMESTIC	-1.540204033
WILD	2.090276902

(4)정준 변수에 의한 산점도

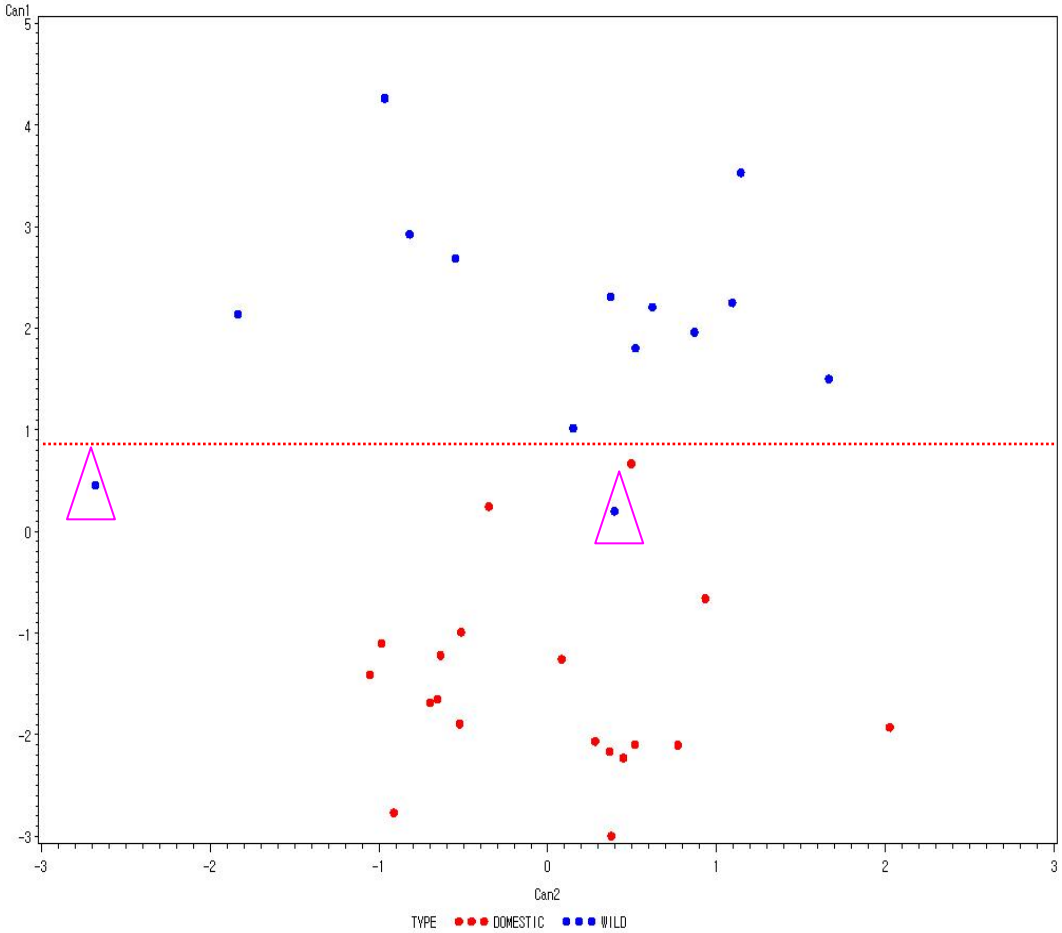
```
PROC PRINT DATA=CANSCORE;
RUN;
```

Obs	ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	Can1	Can2
13	B710	153	140	147	142	151	817	305	102	128	WILD	0.20162	0.39687
14	B790	156	137	151	146	155	814	305	111	137	WILD	2.30958	0.37500
15	B791	.	132	148	138	145	775	.	106	128	WILD	.	.
16	B795	151	134	151	144	.	789	292	116	126	WILD	.	.
17	B819	158	135	151	146	152	790	289	111	125	WILD	2.25104	1.09502
18	B081	.	135	149	.	149	789	.	111	123	WILD	.	.
19	B085	148	129	146	139	147	767	287	106	123	WILD	1.80588	0.52029
20	B089	157	140	154	140	159	818	301	116	136	WILD	4.26364	-0.96511
21	B090	153	138	153	141	151	822	312	115	133	WILD	1.96090	0.87063
22	B091	156	138	156	145	150	835	310	118	133	WILD	1.50190	1.66635
23	B093	151	133	148	139	152	793	290	105	.	WILD	.	.
24	B097	153	135	150	144	158	772	276	102	123	WILD	2.68889	-0.54547
25	B099	152	140	151	144	158	792	303	111	122	WILD	3.53077	1.14522

개체의 집단이 2 개이므로 SYMBOL 문을 2 개 사용하면 된다. 정준 변수(원 변수의 선형 결합)에 의해 개체를 분류할 수 있다.

```
SYMBOL1 V=DOT C=RED;
SYMBOL2 V=DOT C=BLUE;
PROC GPLOT DATA=CANSCORE;
    PLOT CAN1*CAN2=TYPE;
RUN;
```

색을 구별할 수 없으므로 오분류인 점에는 △로 표시하였다. 만약 Bullet 을 다르게 하려면 SYMBOL1 에서 V=DOT 대신 V=CIRCLE 을 하면 선 위 부분의 점은 ○이 된다.



제일 정준에 의해 8 개 변수의 변동 100%가 설명되므로 제이 정준 변수는 개체를 분류하는 역할은 하지 못한다. 제일 정준에 의해 개체를 분류한다면 오분류 2 개만 발생하므로 이것을 판별 분석에 이용할 수 있다. 그러나 이처럼 정준 변수가 하나만 필요한 경우는 극히 이례적인 것이며 2 개 이상 필요하다면 정준 변수에 의해 판별 분석을 실시해야 한다. 비록 제일 정준 변수 하나만으로 충분한 경우라도 새로운 개체를 분류하기 위해서는 프로그램 작업이 필요하다. SAS 가 판별식을 default 로 넣지 않는 것만 보아도 알 수 있듯이 정준 판별 분석은 개체 판단 보다는 표현에 사용되는 도구이다. 다시 정준 변수를 이용하여 판별 분석을 실시한다? 다시 강조하지만 정준 판별 분석은 개체를 저 차원으로 표현하는 방법으로 이용하기 바란다. 그러므로 정준 판별 분석은 실제 다변량 데이터 분석에서 유용성이 떨어진다. 앞의 평균 출력 결과에서 알 수 있듯

이 야생의 CAN1 평균은 2.09 이고 사육 칠면조의 CAN1 평균은 -1.54. 이므로 제일 정준 변수에 의해 칠면조는 확실히 구별될 수 있음을 알 수 있다.

6.6. K Nearest Neighbor 판별 분석

모집단이 정규분포를 따르지 않는 경우 사용하는 비모수 판별 분석 방법으로 개체들의 판별 변수(측정 변수)간의 Mahalanobis 거리($d_i = (\underline{x}_0 - \underline{\mu}_i)'\Sigma^{-1}(\underline{x}_0 - \underline{\mu}_i)$)를 이용하여 개체를 판별하는 방법이다. K nearest neighbor 방법의 절차를 정리하면 다음과 같다.

(1)분류하려는 개체와 Mahalanobis 거리가 가장 가까운 개체를 구하고 그 개체가 속한 집단으로 분류한다.

(2)만약 거리가 같은 개체가 2 개인 경우 동일 집단이면 그 집단에 분류한다.

(3)2 개이면서 그 개체의 집단이 동일하지 않으면 그 다음 가까운 개체의 집단을 조사하여 3 개의 개체 중 많이 속한 집단으로 분류한다. 여기서 k nearest neighbor 의미는 Mahalanobis 거리가 가장 가까운 개체 k 개를 고려하여 그 k 개 개체의 군집 중 가장 많은 수를 차지하는 군집에 분류하게 된다. 다음 프로그램 거리가 가장 가까운 3 개의 개체들의 집단을 조사하여 가장 많은 집단으로 분류하는 방법이다.

6.6.1. 예제

```
PROC DISCRIM DATA=TURKEY METHOD=NPAR K=3 LIST CROSSVALIDATE;
    CLASS TYPE;
    VAR HUM RAD ULN FEMUR TIN CAR D3P COR SCA;
RUN;
```

LIST 옵션은 사후 확률과 분류 결과를 출력하는 명령이고 CROSSVALIDATE 는 오분류 표를 출력하게 하는 옵션이다.

Cross-validation Summary using 3 Nearest Neighbors
Number of Observations and Percent Classified into TYPE

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	19 100.00	0 0.00	19 100.00
WILD	3 21.43	11 78.57	14 100.00
Total	22 66.67	11 33.33	33 100.00

➔ K=3

Cross-validation Summary using 2 Nearest Neighbors

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	18 94.74	1 5.26	19 100.00
WILD	3 21.43	11 78.57	14 100.00
Total	21 63.64	12 36.36	33 100.00

➔ K=2

Cross-validation Summary using 5 Nearest Neighbors

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	19 100.00	0 0.00	19 100.00
WILD	3 21.43	11 78.57	14 100.00
Total	22 66.67	11 33.33	33 100.00

➔ K=5

K=3, K=5 인 경우 오분류 비율이 가장 적고 오분류 형태(부분)도 같으므로 줄 중 어떤 K 를 사용해도 무방하다. TURKEY 자료에서 변수 선택 결과를 이용하여 K=5 nearest neighbor 방법으로 판별 분석을 실시해 보자.

```
PROC DISCRIM DATA=TURKEY METHOD=NPAR K=5 LIST CROSSVALIDATE;
    CLASS TYPE;
    VAR TIN D3P COR ULN;
RUN;
```

Cross-validation Summary using 3 Nearest Neighbors

From TYPE	DOMESTIC	WILD	Total
DOMESTIC	18 94.74	1 5.26	19 100.00
WILD	2 12.50	14 87.50	16 100.00
Total	20 57.14	15 42.86	35 100.00

Error Count Estimates for TYPE

	DOMESTIC	WILD	Total
Rate	0.0526	0.1250	0.0888
Priors	0.5000	0.5000	

변수 선택 후 Fisher 판별 분석 방법(6.4.5 절 참고)을 적용하여 얻은 오분류 개수와 3 개로 동일하지만 오분류 형태가 다르다. Fisher 방법에 의하면 야생을 사육으로 오분류 한 개수는 2 개(10.53%), 반대의 경우는 1 개(6.25%), 오분류 비율은 평균은 8.39%로 K Nearest neighbor 판별 분석의 오분류 비율 8.88%보다 낮으므로 변수 선택 후 Fisher 판별 분석 방법을 이용하는 것이 적절하다.

6.6.2. 새로운 접근 방법

판별 변수(측정 변수)가 이산형, 순서형 분류형, Binary 인 경우 사용되는 Classification Trees 방법이 있다. Breiman, Friedman, Olshen, Stone (1984) 제안한 방법으로 그들의 책 제목은 CART(Classification And Regression Trees)라고 되어 있다. 비슷한 방법으로 J. A. Hartigan 이 개발한 CHAID(Chi-square Automatic Interaction Detector)가 있다. 이 방법은 현재 Data Mining 기법으로 가장 많이 이용되고 있다. SAS E/Minor, SPSS Clementine 등의 Data Mining Tool 에서 제공된다.

6.7. 집단이 3 개 이상인 경우

6.7.1. 예제 자료

밀(Wheat) 종류에는 Arthur 종(soft 한 밀)과 Arkan 종(hard 밀)이 있고 Group 1, 2 과 Group 3, 4 는 서로 다른 지역이다. 그러므로 4 개의 집단이 존재한다. 밀을 제분하여 보면 지역과 종

이 구별되지만 연구자는 밀에 대해 다음 길이를 조사하였다. [**WHEAT.txt**] 밀의 오른쪽(Right) 면에서 면적(Area), 원주(Perimeter), 길이(Length), 폭(breadth) 그리고 아래쪽(down)에서 면적(Area), 원주(Perimeter), 길이(Length), 폭(breadth)을 조사하였다. [Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998, p237]

1	MASO	ARKAN	1	54.4518	219	89	43	56.6039	226	89	47
2	MASO	ARKAN	1	55.1453	221	91	46	56.2583	224	91	46
3	MASO	ARKAN	1	53.9166	223	90	44	55.0908	223	91	44
4	MASO	ARKAN	1	52.2303	212	87	41	53.5444	215	88	44
5	MASO	ARKAN	1	51.5558	207	78	42	52.9811	211	81	44
6	MASO	ARKAN	1	50.4282	203	82	41	51.2152	207	82	42

6.7.2. 판별 분석

변수 집단이 2 개인 경우 동일하게 오분류 표와 사후 확률을 통한 오분류 해석을 하면 된다. 다음은 밀 예제에서 모든 변수를 다 사용하여 Fisher 판별 분석 방법에 의한 분석 결과를 얻었다. 물론 이 판별 방법이 가장 오분류를 적게 하는지는 모른다.

```
DATA WHEAT;
    INFILE 'D:\TEMP\WHEAT.TXT';
    INPUT ID LOC $ TYPE $ GROUP D_A D_P D_L D_B R_A R_P R_L R_B;
RUN;
```

```
PROC DISCRIM DATA=WHEAT CROSSLIST OUTCROSS=OUT1;
    CLASS GROUP;
    VAR D_A D_P D_L D_B R_A R_P R_L R_B;
RUN;
```

```
DATA OUT2;
    SET OUT1;
    IF (GROUP^=_INTO_);
RUN;
```

```
PROC PRINT DATA=OUT2;
    VAR GROUP _INTO_ _1 _2 _3 _4;
RUN;
```

◈오분류 결과

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	3	4	Total
1	24 66.67	1 2.78	11 30.56	0 0.00	36 100.00
2	6 16.67	15 41.67	2 5.56	13 36.11	36 100.00
3	12 24.00	1 2.00	35 70.00	2 4.00	50 100.00
4	1 2.00	13 26.00	6 12.00	30 60.00	50 100.00

◈사후 확률 출력 결과 OUT1 data

Obs	GROUP	_INTO_	_1	_2	_3	_4
10	1	2	0.01659	0.62101	0.00289	0.35951
11	1	3	0.34789	0.10857	0.45931	0.08423
12	1	3	0.01524	0.06085	0.54262	0.38129
13	2	1	0.47332	0.45725	0.01641	0.05301
14	2	1	0.48940	0.04735	0.40174	0.06152

사후 확률이 출력되므로 각 집단에 속할 확률을 알 수 있다. Obs. 10 번은 원래 1 집단이었는데 2 집단의 사후 확률이 0.621 로 가장 크므로 2 집단으로 판별된다. (오분류)

6.7.3. 변수 선택 방법

BACKWARD 변수 선택 방법을 이용하여 유의한 판별 변수를 선택해 보자.

```
PROC STEPDISC DATA=WHEAT METHOD=BACKWARD;
    CLASS GROUP;
    VAR D_A D_P D_L D_B R_A R_P R_L R_B;
RUN;
```

3 개의 변수(D_L, R_P, D_B)가 유의확률이 default 유의수준 0.15 보다 높으므로 판별 변수에서 제외되었다. 유의수준을 높이려면 SLS=0.2 로 하면 된다.

Step	Number In	Removed	Partial R-Square	F Value	Pr > F	Wilks Lambd
0	8		.	.	.	0.2663069
1	7	D_L	0.0146	0.80	0.4979	0.2702559
2	6	R_P	0.0194	1.07	0.3643	0.2756011
3	5	D_B	0.0292	1.63	0.1834	0.2838926

Number of Observations and Percent Classified into GROUP

From GROUP	1	2	3	4	Total
1	26 72.22	1 2.78	9 25.00	0 0.00	36 100.00
2	5 13.89	20 55.56	2 5.56	9 25.00	36 100.00
3	12 24.00	0 0.00	36 72.00	2 4.00	50 100.00
4	1 2.00	17 34.00	4 8.00	28 56.00	50 100.00
Total	44 25.58	38 22.09	51 29.65	39 22.67	172 100.00
Priors	0.25	0.25	0.25	0.25	

각 행의 오분류 개수가 현저히 줄어들어 있음을 알 수 있다. 다음은 오분류 결과를 비교한 것이다.

(1)변수 선택

Error Count Estimates for GROUP

	1	2	3	4	Total
Rate	0.2778	0.4444	0.2800	0.4400	0.3606
Priors	0.2500	0.2500	0.2500	0.2500	

(2)전체 변수 사용

Error Count Estimates for GROUP

	1	2	3	4	Total
Rate	0.3333	0.5833	0.3000	0.4000	0.4042
Priors	0.2500	0.2500	0.2500	0.2500	

K-nearest neighbor 방법도 사용하여 오분류가 가장 적은 판별 분석 방법을 선택한 후 새로운 개체를 분류하면 된다.

6.8. 로지스틱 판별 분석

판별 분석은 판별 변수가 모두 측정형 (연속형: continuous, measurement, metric))인 경우 사용할 수 있다. 물론 decision tree 방법(CART, CHAID)인 경우 판별 변수가 이산형이나 순서형 분류형 변수인 경우도 가능하지만 일반적으로 측정형 변수만이 판별에 이용된다.

로지스틱 회귀 분석(Logistic Regression)은 종속 변수가 이진형(binary, dichotomous: 가질 수 있는 값이 0 또는 1 인 변수)이거나 순서형(ordinal: 상/중/하) 변수인 경우 사용되는 회귀 분석이다. 그러므로 판별 변수가 설명 변수이고 종속 변수가 집단이 된다. 회귀 분석의 변수 선택 방법에 의해 유의한 판별 변수를 선택하면 되고 판별 변수가 측정형 변수가 아니더라도 판별 변수로 사용할 수 있다.

로지스틱 회귀 분석에서 종속 변수 값은 0, 1(사건: 성공, 불량)로 입력된다. 칠면조 예제를 생각해 보자. 야생 칠면조는 경우 $y=1$, 사육 칠면조는 $y=0$ 으로 하여 회귀 분석하면 된다. 로지스틱 회귀 분석에서는 종속변수가 1 혹은 0 을 가질 확률을 추정하게 된다. 추정 확률을 이용하여 개체를 분류하게 된다.

로지스틱 회귀분석은 이진형 반응변수뿐 아니라 반응변수가 순서형(ordinal) 분류형인 경우 사용할 수 있다. 예를 들면 종속 변수가 고객의 신용도이고 이 변수가 (상, 중, 하) 분류되어 있는 경우 사용할 수 있습니다. 종속변수의 수준이 3 개 이상인 경우 LOGISTIC 모형을 사용하는 것이 아니라 CATMOD 를 사용해야 한다고 언급한 책이 있다. 그러나 CATMD 는 CATegorical data MODeling 의 약어로 분류변수 자료 모형화이며 LOGISTIC 모형은 CATMOD 기법의 한 부분입니다

6.8.1 모형

일반 선형 회귀 모형 $y_i = f(\underline{x}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i$, $e_i \sim iidN(0, \sigma^2)$

로지스틱 회귀모형의 종속 변수는 0 과 1 두 값만 가지므로(더 이상 정규분포를 따르지 않는다) 결정계수(R^2)가 매우 낮고 F-검정이나 t-검정을 사용하여 모형, 회귀 계수 추정을 할

수 없다. 만약 종속 변수 y_i가 이진형인 경우(자료가 0, 1 만 존재) OLS 에 의한 계수 추정은 무의미 하다.

(1)ODDS: $odd = \frac{p}{1-p}$

어떤 사건이 발생할 가능성[p=0.5 일 경우 1 이다. 기준]으로 해석될 수 있다. 한국이 2002 년 16 강에 들어갈 확률 0.1 이면 1/9 이 Odds 이다. ▶ 1$ betting 에서 이기면 9$ return 브라질이 2002 년 16 강에 들어갈 확률 0.8 이면 4 가 Odds 이다. ▶ 4$ betting 에서 이기면 1$ return

(2)ODDS transformation (변환)

$$p^* = \frac{p}{1-p}$$

(3)로지스틱 회귀 모형

종속 변수를 $p_i = \Pr(Y=1)$ 라고 생각해 보면 종속 변수는 어떤 사건이 일어날 확률이 $(Y=1)$ 된다. 여기에 odds 개념을 적용하여 Odds 변환을 해 보자.

$$p_i^* = \frac{p_i}{1-p_i}$$

확률 p_i가 (0,1) 사이의 값을 가지므로 p_i^*는 (0, ∞) 값을 가진다. $\ln(p_i^*)$ 변환을 하면 이 변수는 (-∞,∞) 값을 가지므로 아래 모형에서 오차항의 $e_i \sim Normal(0,\sigma^2)$ (회귀 분석 가정)에는 문제가 없을 것이다. 이 모형을 Logistic model 이라 한다.

$$y_i = \ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i \,,\ \ e_i \sim Normal(0,\sigma^2)$$: 로지스틱 모형

위의 모형을 다시 쓰면 다음과 같다.

$$p_i = \Pr(Y=1 \mid \underline{x}) = \frac{e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}}{1+e^{\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i = \frac{1}{1+e^{-\{\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+..+\beta_p x_{pi}\}}} + e_i$$

그러므로 회귀 계수의 부호가 양수이고 값이 커지면 p_i (성공: $Y=1$)가 커지므로 성공 확률이 높아지고 부호가 음수이고 절대값이 커지면 p_i 가 작아지므로 성공 확률이 낮아진다.

(4)모형의 적합성 검정 및 회귀계수 유의성 검정

모형 전체의 유의성은 –2Log L, AIC(Akaike Information Criterion) Schwartz Criterion 을 이용하고 (Adjusted 결정계수와 유사 개념) 회귀계수의 유의성 검정은 Wald 의 Chi-square 검정통계량을 이용한다.

6.8.2. OLS 추정 문제점

II EXAMPLE **II**

Remission.txt 자료는 환자의 상태를 나타내는 변수 (cell, smear, infil, li, blast, temp)들이 암 재발 여부(종속변수)에 영향을 미치는지 알아보기 위하여 수집한 자료이다.

```
data Remission;
     input remiss cell smear infil li blast temp;
     datalines;
  1   .8   .83   .66  1.9  1.1     .996
  1   .9   .36   .32  1.4   .74    .992
  0   .8   .88   .7    .8   .176   .982
```

remiss=1 이면 재발

remiss=0 이면 재발하지 않음.

OLS 의 문제점을 잘 파악하기 위하여 설명 변수가 하나(Li 만)인 경우 살펴보기로 한다. 우선 OLS 방법에 의해 회귀 분석을 실시해 보자.

```
proc reg data=Remission;
    model remiss=li;
    title h=1.5 'scatter plot of remisss vs. li';
    plot remiss*li;
    title h=1.5 'residual plot';
    plot residual.*predicted.;
    output out=out1 p=yhat_o;
run;
```

OLS 추정치에 의한 $\hat{Y}$ 결과 OUT1 에 저장

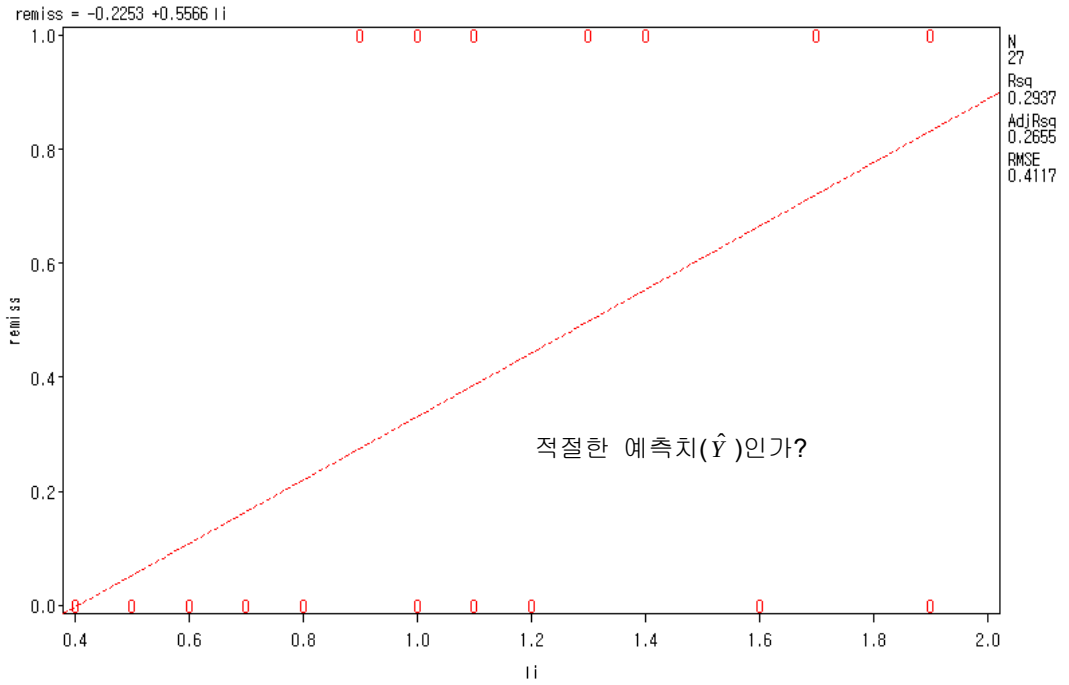
Root MSE	0.41171	R-Square	0.2937
Dependent Mean	0.33333	Adj R-Sq	0.2655
Coeff Var	123.51163		

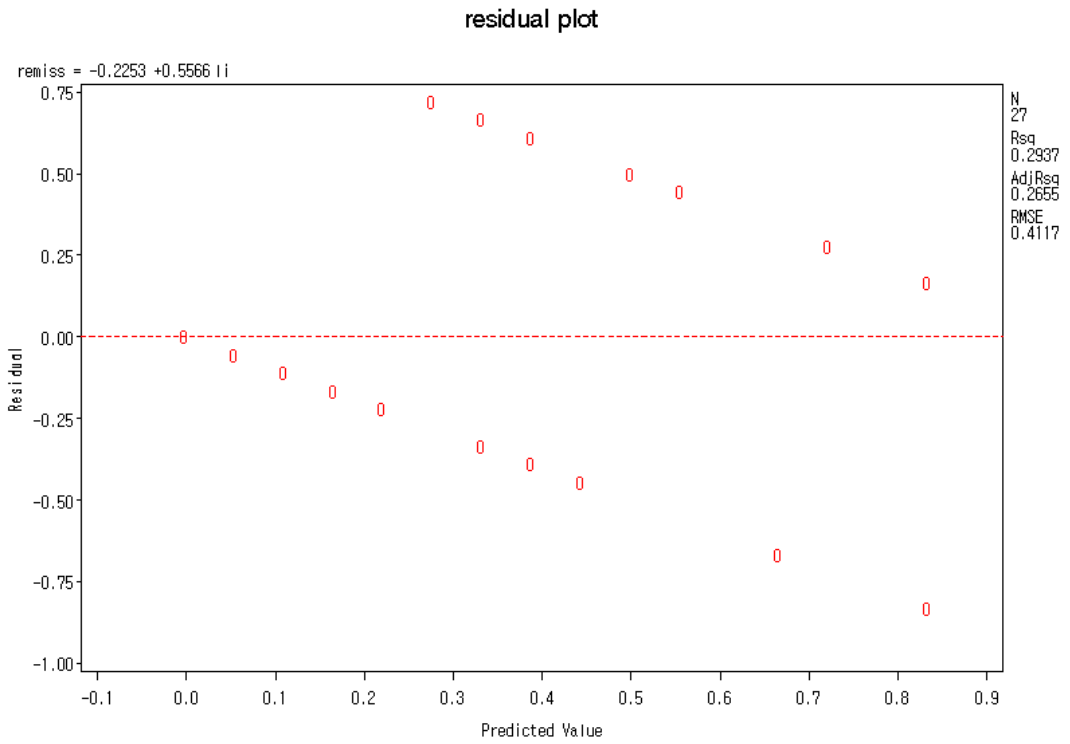
Li 회귀 계수 유의성은 매우 유의하지만 (p=0.0035) 결정 계수가 매우 낮다.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.22530	0.19050	-1.18	0.2481
li	1	0.55657	0.17260	3.22	0.0035

scatter plot of remisss vs. li





6.8.3. 로지스틱 회귀 분석

OLS 분석 방법과 비교하기 위하여 설명 변수가 Li 하나인 경우 로지스틱 분석을 실시해 보자.

```
proc logistic descending data=Remission;
     model remiss=li;
     output out=out2 p=yhat_l;
run;
```

descending 옵션을 사용하는 이유? SAS 는 코딩 값이 작은 것을 event(사건)라 보고 큰 것을 non-event 라 본다. 그런데 예제 자료는 1 이 재발(이것이 event 에 해당)이므로 자료 코딩을 반대로 인식하라는 명령으로 descending 을 사용한다. output 문장에 의해 로지스틱 회귀 모형 추정에 의한 예측치($\hat{Y}$) 결과를 OUT2 에 저장했다.

Response Profile

Ordered Value	remiss	Total Frequency	
1	1	9	← event
2	0	18	← non-event

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.2988	1	0.0040
Score	7.9311	1	0.0049
Wald	5.9594	1	0.0146

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.7771	1.3786	7.5064	0.0061
li	1	2.8973	1.1868	5.9594	0.0146

로지스틱 회귀 분석에서는 회귀 계수의 유의성 검정은 χ^2-검정에 의한다. p-값이 0,0146 이므로 Li 설명 변수는 매우 유의하다.

```
data out;
    merge out1 out2;
run;
```

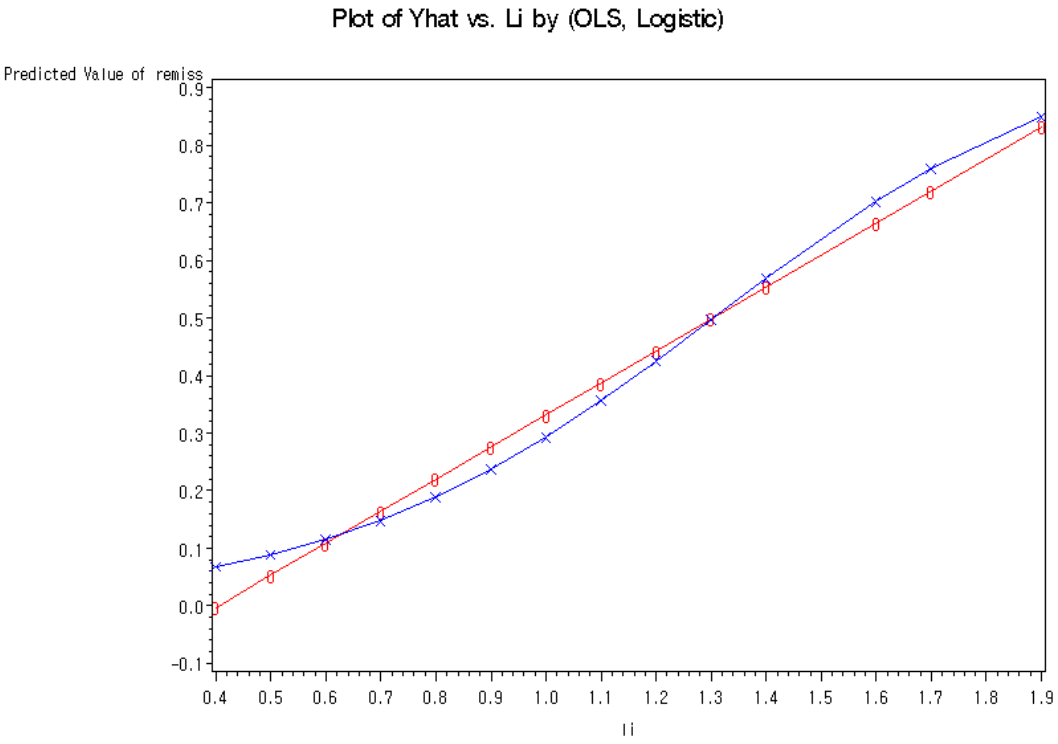
OLS 에 의한 $\hat{Y}$ ($Yhat\_o$), 로지스틱에 의한 $\hat{Y}$ ($Yhat\_l$) 합치기

```
proc sort data=out;
    by li;
run;
```

데이터를 Li 값으로 정렬하여야 예측치를 이으면(i=join) 직선을 이나 곡선을 얻는다.

```
proc gplot data=out;
    title h=1.5 'Plot of Yhat vs. Li by (OLS, Logistic)';
    symbol1 v='o' c=red i=join;
    symbol2 v='x' c=blue i=join;
    plot (yhat_o yhat_l)*li/overlay;
run;
```

앞에서 언급하였듯이 OLS 추정치(빨간 선)는 확률이 음으로 예측되기도 한다. 종속변수가 0 과 1 인 경우 회귀 모형에 의한 예측치는 $\hat{Y}_i = \hat{p}_i = \Pr(Y=1|\underline{x})$, 즉 event(본 예제에서는 재발)이 일어날 확률이므로 0 과 1 사이 값이어야 한다.



6.8.3. 로지스틱 이용하여 판별하기

칠면조 예제를 사용해 Logistic 회귀분석을 실시해 보자. 모형 설정, 변수 선택 방법 등은 회귀 분석과 동일하고 분류표(Cross Table)가 출력되는 것만 상이하다.

```
PROC LOGISTIC DATA=TURKEY;
    MODEL TYPE=HUM--SCA/SELECTION=STEPWISE SLE=0.2 SLS=0.1;
RUN;
```

로지스틱 회귀 분석에서도 변수 선택 방법이 가능하다. 사용한 방법은 stepwise 이고 Entry 유의 수준은 0.2, Stay 유의 수준은 0.1 로 하였다. (유의 수준은 SLE=0.25 이상, SLS=0.15 정도가 일반적)

Response Profile

Ordered Value	TYPE	Total Frequency
1	DOMESTIC	19
2	WILD	14

SAS 는 Order Value=1 을 성공(event) 집단, Order Value=2 는 실패(non-event) 집단으로 간주한다. 이 예제에서는 사육(domestic) 칠면조가 성공 집단임을 나타낸다.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	TIN		1	1	21.6200	.	<.0001
2	FEMUR		1	2	5.6931	.	0.0170
3		FEMUR	1	1	.	1.8810	0.1702

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	73.3164	27.1245	7.3060	0.0069
TIN	1	-0.5027	0.1863	7.2855	0.0070

추정 모형은 $p_i = \Pr(Y = Domestic \mid \underline{x}) = \frac{1}{1+e^{-\{73.3-0.5*TIN\}}}$ 이다. TIN 의 값이 작아지면 칠면조가 사육 칠면조일 가능성 높아진다.

(1)INCLUDE 옵션 사용하기

개체들을 분류하기 위하여 측정된 변수 중 유의 수준에 상관 없이 꼭 사용하고픈 변수를 판별에 이용하고자 할 때 INCLUDE 문장을 사용하면 된다. TURKEY 예제에서 Fisher 판별 분석에 의한 변수 선택을 보면 TIN➔D3P➔COR➔ULN 순서이므로 TIN 과 D3P 를 반드시 사용하여 Logistic 회귀 분석의 변수 선택을 실시해 보자. (TURKEY 예제의 경우 변수 선택이 잘되지 않으므로 보여 주기 위하여 유의 수준을 높여 보자) INCLUDE=2 옵션은 설명 변수 중 앞의 2 개 변수를 반드시 선택하게 한다. 즉 TIN, D3P 변수는 절대 제외되지 않는다.

```
PROC LOGISTIC DATA=TURKEY ;
    MODEL TYPE=TIN D3P HUM RAD ULN FEMUR CAR COR SCA
    /SELECTION=STEPWISE SLE=0.5 SLS=0.4 INCLUDE=2;
RUN;
```

Summary of Stepwise Selection

Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
FEMUR		1	3	6.3280	.	0.0119
COR		1	4	0.7154	.	0.3977
	COR	1	3	.	0.4794	0.4887

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	544.2	513.9	1.1216	0.2896
TIN	1	-7.5548	7.0104	1.1613	0.2812
D3P	1	0.6701	0.7257	0.8526	0.3558
FEMUR	1	2.6159	2.5876	1.0220	0.3120

각 변수의 p-값은 매우 높음을 알 수 있다.

Tentative 추정 모형은 $p_i = \Pr(Y = Domestic \mid \underline{x}) = \frac{1}{1+e^{-\{544.2-7.55*TIN+0.67*D3P+2.61*FEMUR\}}}$

이다. TIN 의 값이 작아질수록, D3P 와 FEMUR 값은 커질수록 칠면조가 사육 칠면조일 가능성 높아진다.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.3516	3	<.0001
Score	24.9784	3	<.0001
Wald	1.1932	3	0.7546

귀무가설(3 개 변수가 유의하지 않다. 3 개 변수의 회귀 계수는 모두 0 이다)을 검정하는 검정통계량 (결론: 적어도 하나 이상은 유의)

(2)교차표와 표준화 회귀 계수(Standardized Beta Coefficient)

CTABLE 옵션은 로짓 판별 분석을 통한 개체 판별 결과에 대한 교차표(cross-tabulation)를 작성한다. STB 옵션은 표준화 회귀 계수(설명 변수의 영향 정도를 비교할 수 있다)를 추정하게 하는 옵션이다. (두 변수 TIN, D3P 만 사용하였다). 우선 표준화 회귀 계수를 해석해 보자.

```
PROC LOGISTIC DATA=TURKEY ;
    MODEL TYPE=TIN D3P /CTABLE STB;
RUN;
```

Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	64.4169	25.9556	6.1594	0.0131	
TIN	-0.9240	0.3813	5.8737	0.0154	-4.3783
D3P	0.2343	0.1280	3.3523	0.0671	1.5423

변수 TIN 의 표준화 회귀 계수가 –4.37 로 1.54 보다 절대 값이 크므로 칠면조가 사육(Event)일 확률에 더 많은 영향을 미친다. 측정 단위가 다른 경우 일반 회귀 계수(-0.92, 0.23)를 이용하여 변수의 영향 정도를 평가해서는 안 된다.

CTABLE 옵션은 로짓 판별 분석을 통한 개체 판별 결과에 대한 교차표(cross-tabulation)를 작성한다. STB 옵션은 표준화 회귀 계수를 추정하게 하는 옵션이다. (두 변수 TIN, D3P 만 사용하였다).

Ordered Value	TYPE	Total Frequency
1	DOMESTIC	19
2	WILD	18

위에 있는 집단이(DOMESTIC) EVENT 로 정의되므로 우리가 성공(값=1)이라 생각되는 것이다. 즉 Pr(Y=1)의 의미는 Pr(Y=Event)와 같다.

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	19	0	18	0	51.4	100.0	0.0	48.6	.
0.020	19	11	7	0	81.1	100.0	61.1	26.9	0.0
0.040	19	13	5	0	86.5	100.0	72.2	20.8	0.0
0.060	19	14	4	0	89.2	100.0	77.8	17.4	0.0
0.080	19	14	4	0	89.2	100.0	77.8	17.4	0.0
0.100	19	14	4	0	89.2	100.0	77.8	17.4	0.0
0.120	18	14	4	1	86.5	94.7	77.8	18.2	6.7
0.140	18	14	4	1	86.5	94.7	77.8	18.2	6.7
0.160	18	15	3	1	89.2	94.7	83.3	14.3	6.3
0.180	18	15	3	1	89.2	94.7	83.3	14.3	6.3
0.200	17	16	2	2	89.2	89.5	88.9	10.5	11.1
0.220	17	16	2	2	89.2	89.5	88.9	10.5	11.1
0.240	17	17	1	2	91.9	89.5	94.4	5.6	10.5
0.260	17	17	1	2	91.9	89.5	94.4	5.6	10.5
0.280	17	17	1	2	91.9	89.5	94.4	5.6	10.5
0.300	17	17	1	2	91.9	89.5	94.4	5.6	10.5
0.320	17	17	1	2	91.9	89.5	94.4	5.6	10.5
0.340	17	17	1	2	91.9	89.5	94.4	5.6	10.5

①Prob. Level 은 Logit 회귀분석에 의해 추정된 Pr(Event) 확률이 이 값 이상이면 Event 로 정의한다는 의미이다. 첫 열을 보면 추정된 Pr(Event) 확률이 0.0 이상이면 Event 로 분류한다면 가정하고 개체를 분류한 결과를 보여준다. 그러므로 EVENT 19 개는 모두 Event 로 정분류 되고 non-EVENT 18 개는 모두 Event 로 분류되므로 오분류 된다.

②Correct Event(Domestic) 칠면조를 Event 로 Non-event 칠면조를 Non-event 로 정분류

③In-Correct Event(Domestic) 칠면조를 non-Event 로 Non-event 칠면조를 event 로 오분류

④Correct 는 전체 개체 수 중 정분류 된 개체 비율을 출력한 것이다. 첫 열의 51.4 는 19/(19+18)이다. 두번째 열의 81.1 은 (19+7)/(19+18)... 이렇게 계산된다.

⑤Sensitivity 는 Event(사육) 개체를 Event(사육)으로 정분류 비율

⑥False Pos. 는 Event(사육) 개체를 non-Event(야생)으로 오분류 비율, 그러므로 Sensitivity + False Pos. 는 1 이다.

⑦Specificity 는 non-Event(야생) 개체를 non-Event(야생)으로 정분류 비율

⑧False Neg. 는 non-Event(야생) 개체를 Event(사육)으로 오분류 비율, 그러므로 Specificity + False Neg. 는 1 이다.

CORECT 가 가장 적은 Prob. Level 기준 값을 이용하여 개체를 분류하면 된다. 물론 Sensitivity 와 Specificity 도 함께 고려해야 한다. (오분류 형태를 보여 주는 것이므로) 칠면조 자료 예제의 경우 Prob. Level 이 0.24~0.8 사이의 어떤 값을 사용하든 Correct, Sensitivity, Specificity 비율이 동일하다. 나는 Prob. Level 을 0.4 로 하였다. 그러므로 Pr(Event)의 추정 확률이 0.4 이상이면 Event(사육)으로 분류하고 미만이면 non-Event(야생)로 분류한다. 0.4 를 사용하면 오분류 개수는 3 개(오분류 비율 8.1%=3/37, Non-event 을 Event 로 Event 를 Non-event 로 분류하였다) 앞의 Fisher 방법이나 K-nearest 방법과 오분류가 동일하다.

6.8.4. 새로운 개체 판별 분석

다음 프로그램은 Pr(Event)의 추정치를 Phat 라는 이름으로 하여 SAS data OUT1 에 저장하였다.

```
PROC LOGISTIC DATA=TURKEY ;
     MODEL TYPE=TIN D3P /CTABLE STB;
     OUTPUT OUT=OUT1 P=PHAT;
RUN;
PROC PRINT DATA=OUT1;
RUN;
```

Obs	ID	HL	SCA	TYPE	_LEVEL_	PHAT
29	B105		134	WILD	DOMESTIC	.
30	B106	1	131	WILD	DOMESTIC	0.13614
31	B111	1	128	WILD	DOMESTIC	0.04119
32	B114	1	132	WILD	DOMESTIC	0.00001
33	B116	1	134	WILD	DOMESTIC	0.01015
34	B117		136	WILD	DOMESTIC	0.00005

_LEVEL_은 무조건 EVENT(Domestic)만으로 분류해 놓는다. 그러므로 앞에서 정한 Prob. Level 기준 값을 0.4 로 하였으므로 다음에 의해 분류하면 된다. PHAT 값이 0.4 미만이면 WILD(야생)으로 분류하여 _LEVEL_ 에 WILD 로 바꾼다. 다음 라인은 PHAT 값이 존재하는 것만 출력하게 한다. 만약 판별에 사용된 측정 변수에 결측치가 없으면 이 문장은 필요 없다.

```
DATA OUT2;
    SET OUT1;
         IF (PHAT<0.4) THEN _LEVEL_="WILD";
         IF (PHAT>=0);
RUN;

PROC PRINT DATA=OUT2;
    VAR TIN D3P TYPE _LEVEL_ PHAT;
RUN;
```

Obs	TIN	D3P	TYPE	_LEVEL_	PHAT
16	151	301	WILD	WILD	0.01015
17	156	298	WILD	WILD	0.00005
18	141	290	WILD	DOMESTIC	0.88921
19	131	250	DOMESTIC	DOMESTIC	0.87546
20	141	290	DOMESTIC	DOMESTIC	0.88921
21	135	300	DOMESTIC	DOMESTIC	0.99995
22	136	300	DOMESTIC	DOMESTIC	0.99988
23	146	310	DOMESTIC	DOMESTIC	0.89557
24	139	300	DOMESTIC	DOMESTIC	0.99812
25	137	300	DOMESTIC	DOMESTIC	0.99970
26	136	300	DOMESTIC	DOMESTIC	0.99988
27	136	280	DOMESTIC	DOMESTIC	0.98738
28	145	310	DOMESTIC	DOMESTIC	0.95576
29	147	300	DOMESTIC	WILD	0.24634
30	131	310	DOMESTIC	DOMESTIC	1.00000

오분류 결과를 보면 앞의 CTABLE 결과와는 달리 2 개 밖에 없다. 이상하다. 왜 이렇지. 새로운 개체(칠면조) 2 개를 logistic 판별 분석(판별 변수 2 개 사용: TIN, D3P)에 의해 분류해 보자.

```
DATA TEMP;
    TIN=145; D3P=320; OUTPUT;
    TIN=150; D3P=300; OUTPUT;
RUN;

DATA TURKEY;
    SET TURKEY TEMP;
RUN;
```

```
PROC LOGISTIC DATA=TURKEY ;
    MODEL TYPE=TIN D3P/CTABLE STB;
    OUTPUT OUT=OUT1 P=PHAT;
RUN;
PROC PRINT DATA=OUT1;
    VAR TIN D3P TYPE _LEVEL_ PHAT;
RUN;
```

원래 자료 마지막에 새로운 개체 2 개에 대한 Pr(EVENT)가 출력된다. 앞에서 기준 값을 0.4 로 하였으므로 (TIN=145, D3P=320) 칠면조는 Domestic(사육)으로 (TIN=150, D3P=300)인 칠면조는 Wild (야생)으로 분류하면 된다.

Obs	TIN	D3P	TYPE	_LEVEL_	PHAT
82	136	.	DOMESTIC	DOMESTIC	.
83	145	320		DOMESTIC	0.99558
84	150	300		DOMESTIC	0.02003

6.8.5. 로지스틱 판별 분석의 장점

Fisher 판별 분석과 K nearest neighbor 판별 분석 방법은 측정 변수가 모두 측정형이어야 한다. (물론 CART, CHAID 방법은 분류형 변수도 가능하지만) 그러나 Logistic 판별 분석은 1)설명 변수(판별 변수)가 분류형이어도 가능하고 2)설명 변수(판별 변수)의 영향 정도를 비교할 수 있다. 우리는 개체가 2 개인 경우(이진형 변수)만 살펴 보았는데 순서형 변수인 경우에도 Logistic 판별 분석이 가능하다. 또한 판별 변수와 반응 변수(집단) 간의 인과 관계와 인과 관계 정도(영향력: 표준화 회귀 계수)를 알 수 있다는 장점이 있다.

개체의 집단이 3 개 이상이고 순서형인 경우 Logistic 회귀 분석 결과 해석 방법을 살펴 보자. WHEAT 예제에서 GROUP 을 순서형 분류형이라 가정 하자(**1<2<3<4 의 순서가 있다면 ……**).

```
PROC LOGISTIC DATA=WHEAT;
    MODEL GROUP= D_A D_P D_L D_B R_A R_P R_L R_B
    /SELECTION=STEPWISE SLS=0.2 SLE=0.1 CTABLE;
    OUTPUT OUT=OUT1 P=YHAT;
RUN;

PROC PRINT DATA=OUT1;
RUN;
```

Response Profile

Ordered Value	GROUP	Total Frequency
1	1	36
2	2	36
3	3	50
4	4	50

➔ 개체 집단에 대한 정보, 4 집단이다.

Logistic 모형: $\Pr(Y \le k) = \frac{1}{1+e^{-\{\alpha_i+\beta_1*x_1+\beta_2*x_2+...+\beta_p*x_p+e_i\}}}, k = 1,2,...,g$ g =집단 수

집단이 4 개이므로 그룹 절편은 3 개가 출력되었다. $\alpha_1 = 0, \alpha_2 = 18.16, \alpha_3 = 19.47, \alpha_4 = 21.03$

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	18.1592	4.1856	18.8223	<.0001
Intercept2	1	19.4651	4.2167	21.3096	<.0001
Intercept3	1	21.0364	4.2608	24.3765	<.0001
D_A	1	-0.6500	0.1035	39.4055	<.0001
R_A	1	0.2223	0.1341	2.7503	0.0972
R_P	1	0.1236	0.0440	7.8701	0.0050
R_L	1	-0.2630	0.0619	18.0792	<.0001

개체 1(Y_1)에 대해 3 개의 Phat 결과가 출력된다.

- $\Pr(Y_1 = 1 \mid \underline{x} = (da = 54, ra = 56, rp =, 226, rl = 89)) = 0.467$
- $\Pr(Y_1 = 2 \mid \underline{x} = (da = 54, ra = 56, rp =, 226, rl = 89)) = 0.764 - 0.467 = 0.297$
- $\Pr(Y_1 = 3 \mid \underline{x} = (da = 54, ra = 56, rp =, 226, rl = 89)) = 0.939 - 0.764 = 0.175$
- $\Pr(Y_1 = 4 \mid \underline{x} = (da = 54, ra = 56, rp =, 226, rl = 89)) = 1 - 0.939 = 0.06$

Obs	ID	LOC	TYPE	GROUP	_LEVEL_	YHAT
1	1	MASO	ARKAN	1	1	0.46753
2	1	MASO	ARKAN	1	2	0.76421
3	1	MASO	ARKAN	1	3	0.93975
4	2	MASO	ARKAN	1	1	0.19298
5	2	MASO	ARKAN	1	2	0.46884
6	2	MASO	ARKAN	1	3	0.80946

(누적 확률이 출력된다)

[EXERCISE]

▪[CAR.TXT 데이터 출처: http:// lib.stat.cmu.edu/DASL]

(1)CAR.TXT 데이터에서 5개 (MPG—DISPLACEMENT) 변수를 이용하여 차를 2개 집단(군집)으로 분류하고 개체 그룹(군집)에 적절한 이름을 붙이시오. 군집 분석을 이용하지 말고 주성분 분석을 이용하시오. 주성분이 2개이면 산점도, 한 개이면 줄기-잎 그림을 그리고 개체를 그룹화 하시오. 주성분이 3개 이상이면? 2개만 이용하시오.

(2)자동차를 (US, non-US)로 분류할 때 HOMEWORK #7-1에서 얻은 자신의 판별 규칙에 의해 자동차를 분류한다고 가정하고(물론 이것은 제대로 된 판별 분석은 아니다) 분류표를 작성하시오.

(3)CAR.TXT 데이터에서 2개 (MPG, HORSEPOWER) 변수를 이용하여 자동차의 (US, non-US) 생산국을 판별하는 식을 만들려고 한다.

①생산국에 따라 MPG, HORSEPOWER 산점도를 그리시오.

②FISHER 판별 방법에 의해 자동차의 생산국을 판별하고 CROSS-VALIDATION 방법에 의해 오분류 표를 작성하시오.

③(1)에서 여러분이 정한 판별 규칙과 비교하고 해석하시오.

(4)새로운 차가 2 종류 출시되었는데 생산 국가를 알 수 없다. (3)에서 얻은 판별식을 이용하여 생산국가를 분류하시오.

(MPG, HORSEPOWER)=(20, 100) (MPG, HORSEPOWER)=(25, 120)

(5)CAR.TXT 자료, 판별 변수 2개 MPG, HORSEPOWER, 집단 (US, non-US)이 2개일 때

①미국, non-미국 차의 비율이 70%, 30%라는 사전 정보가 있다. 이를 이용하여 개체를 판별하고 오분류 표를 얻으시오.

②등분산 가정이 검정하시오.

(6)CAR.TXT 자료, 판별 변수 (MPG—DISPLACEMENT) 5개, 집단 (US, non-US) 일 때 PROC STEPDISC 사용하지 말고 판별에 유의한 변수를 구해 보자. (분산 분석 &공분산 분석) 유의 수준=0.2 (Type III SS를 이용하여 Group의 유의성을 판단해야 합니다.)

(7)CAR.TXT 자료, 판별 변수 (MPG—DISPLACEMENT) 5개, 집단 2개(US, non-US)

①5개 변수 모두 사용하여 개체를 분류하시오.

②유의한 판별 변수를 선택하고 개체를 분류하시오.

③MPG, MANPOWER 2개 변수 사용한 결과, ①과 ②의 오분류 결과를 비교 해석하시오.

(8)CAR.TXT에서 5개 측정 변수(MPG—DISPLACEMENT)에 대해 정준 판별 분석을 실시하고 해석하시오.

(9)CAR.TXT에서 5개 측정 변수(MPG—DISPLACEMENT)를 이용하여 K nearest 판별 분석을 사용할 경우 오분류가 가장 적은 K를 구하시오.

(10)Fisher 판별 방법, K-nearest 판별 분석 방법, 변수 선택 방법 중 가장 좋은 판별 분석 방법을 결정하시오.

(11)밀(Wheat) 예제 [WHEAT.txt] 4개 집단, 8개 변수

①Fisher 판별 분석 방법에 의해 오분류 표를 얻으시오. (8개 변수 모두 사용)

②Fisher 판별 분석 방법에서 변수 선택 방법을 이용하여 오분류 표를 계산하시오.

③K-nearest 방법에 의해 오분류 표를 작성하시오. 가장 적합한 K을 얻으시오. (②에서 판별 능력이 있는 변수만 이용하시오.)

④Location집단으로 하여 로지스틱 회귀 분석을 실시하고 (변수 선택 방법을 이용하시오) 장 적절한 cut-off 확률을 정하시오.

⑤①-③ 방법 중 가장 적합한 방법을 제시하고 최종 판별 분석 방법에 의해 오분류가 발생한 개체에 대해 왜 그런지 해석하시오.(평균 비교) 새로운 개체를 임의로 2개 만들어 판별해 보시오,