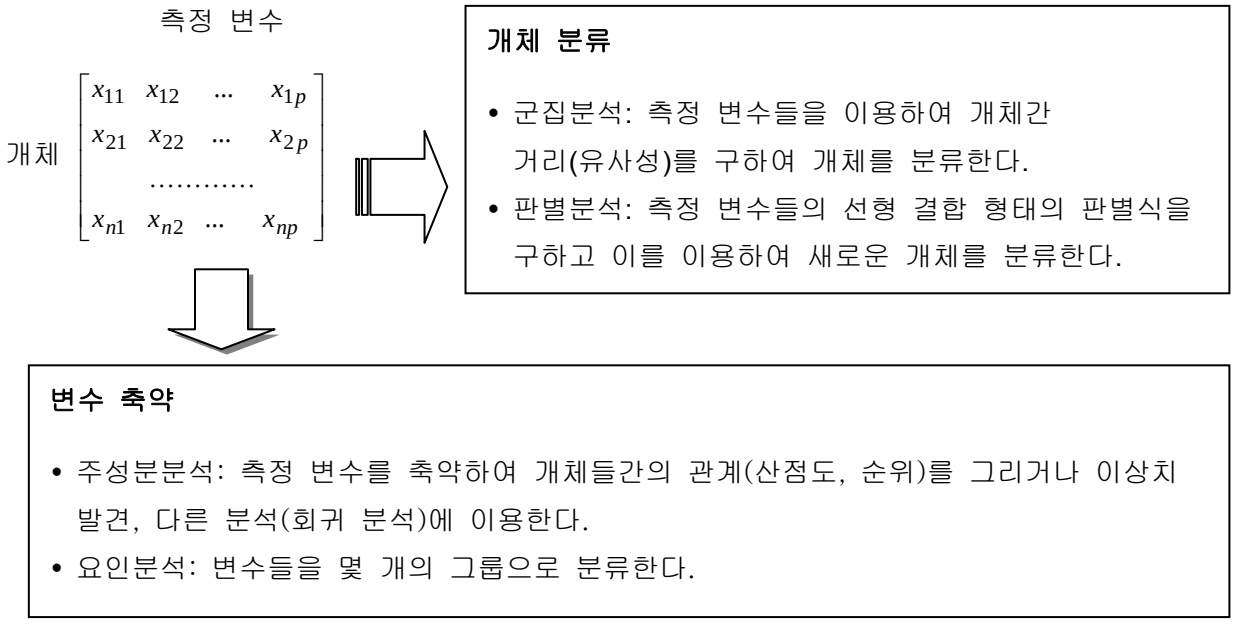
Chapter 4. 판별분석

주성분분석과 요인분석은 $x_1, x_2, ..., x_p$ 변수들의 상관 관계(공분산행렬/상관행렬)를 이용하여 변수 차원을 축약하거나 변수를 그룹화하는 방법으로 변수유도기법(Variable-directed techniques)이라 한다. 판별분석(Discriminant Analysis)은 군집분석(Clustering Analysis)과 함께 개체들에 대해 측정된 특성(변수) 관측치를 이용하여 개체 분류하는 방법으로 개체유도기법(individual directed techniques)라 일컫는다.



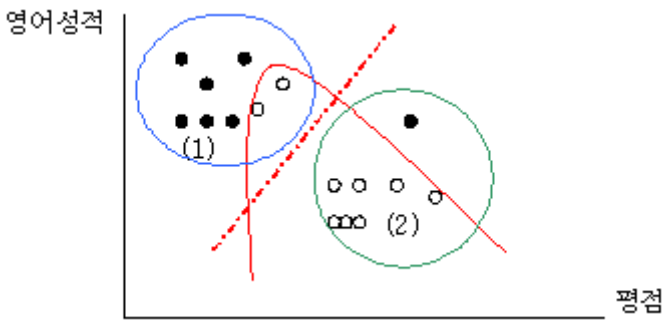
4.1 판별분석 개념

4.1.1 군집분석과 판별분석

대학 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 군집분석이라 한다. 개체 그룹의 수는 분석자가 임의로 정하게 되며 그룹 내의 개체를 본 후 그 그룹의 이름을 부여한다.

판별분석은 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 판별분석이라 한다. 이처럼 군집분석과 판별분석의 차이는 그룹이 분석 전에 알려져 있는가 아닌가 하는 것이다.

16 명(남자 9 명/여자 7 명)의 영어성적과 평점을 조사해 다음 산점도를 얻었다고 하자.



군집분석에서는 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 주성분분석을 이용하여 변수를 축약하고 주성분 변수를 이용하여 군집분석을 실시하면(산점도를 그릴 수 있다) 된다. 주성분분석을 이용한 군집분석은 주성분분석 사용 예에서 중요하다.

판별분석은 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. 판별식(2)를 사용하면 오분류가 없으나 이 곡선 식을 구하는 것은 불가능하다. 그러므로 구하기 쉬운 직선 형태의 판별식(Fisher 의 Linear Deterministic Function) (1)을 이용하게 된다.

4.1.2 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 의료 사고로 이어질 수 있으므로 오분류②에 의해 발생하는 비용이 훨씬 크다. 일반적으로 오분류 비용 계산이 어려우므로 오분류 비용은 동일하다고 가정한다. 그러므로 판별분석은 오분류를 최소화 할 수 있는 판별식을 구하는 것이 주목적이다.

4.1.3 판별 규칙

설명의 간편을 위하여 모집단이 두 개인 경우를 생각하자. 다변량 정규분포를 따르는 2 개의 모집단이 있다고 가정하자.

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

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 함수는 같다고 가정한다.

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.

개체를 판별하는 판별규칙을 정리하면 다음과 같다.

(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. 이것을 Fisher 의 판별식이라고도 한다.

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

(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.

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

(4)사후 확률(Posterior Probability) 규칙

모집단이 동일한 분산-공분산 행렬을 가질 때 모집단 π_1의 사후 확률을 다음과 같이 정의하고 만약 $P(\pi_1 | \underline{x}_0) > P(\pi_2 | \underline{x}_0)$ 이면 π_1 으로 분류하고, 그렇지 않으면 π_2 으로 분류한다.

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

4.1.4 비용함수와 사전확률

모집단이 2 개인 경우 오분류 비용함수를 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 사전비율을 p_1, 2 집단의 사전비율을 p_2라 하자. 집단 평균으로부터 개체의 거리는 다음과 같이 정의한다.

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

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

$d_1^* > d_2^*$이면 1 집단(π_1)으로 분류되고 $d_1^* < d_2^*$이면 2 집단(π_2)으로 분류된다.

현실에서는 비용함수를 고려하는 것은 매우 어려우므로 비용함수는 동일하다는 가정을 하게 된다. 사전 확률을 옵션을 사용하는 경우는 판별에 사용되는 데이터의 집단 구성 비율이 모집단의 비율과 현저히 다르고 모집단 비율을 알고 있을 때이다.

SPSS 는 사전확률을 설정할 수 있는 옵션은 없고 동일하게 설정하거나 표본의 크기 비율을 사용하는 옵션만 제공하고 있다. 비용함수에 대한 옵션은 없다.

4.1.5 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. SPSS 는 비용함수를 고려하는 옵션이 없다.

4.2 판별분석하기

4.2.1 변수 2 개인 경우

미국 Kansas 주립대학 Dr. Michael Finnegan 교수는 야생 칠면조와 사육 칠면조를 구별하기 위하여 수컷 칠면조 82 마리에 대해 9 개 항목을 조사하였다. ■TURKEY.SAV■

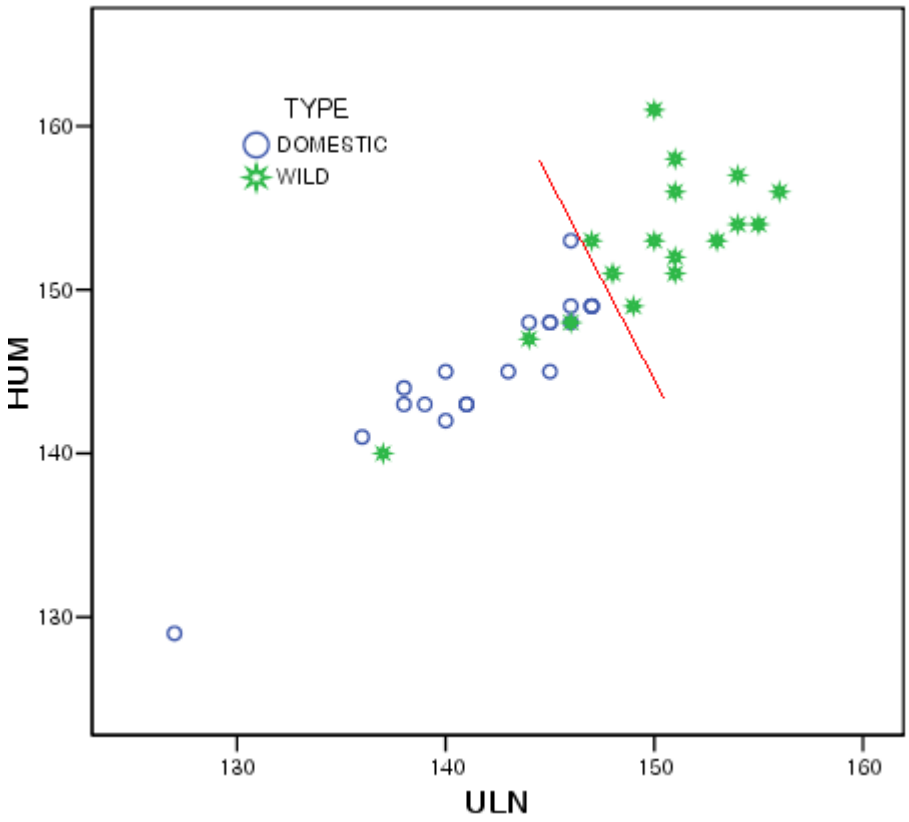
[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998]

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)	

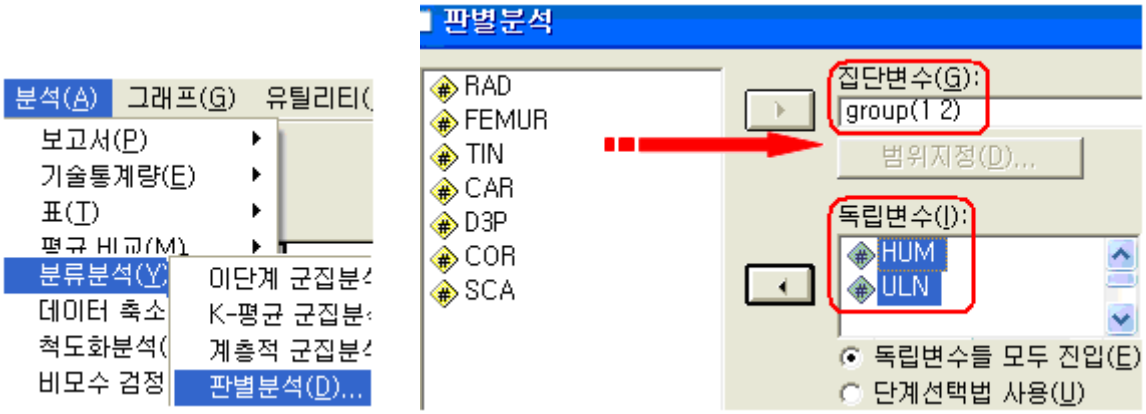
ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE
B710	153	140	147	142	151	817	305	102	128	WILD
B790	156	137	151	146	155	814	305	111	137	WILD
B791	.	132	148	138	145	775	.	106	128	WILD
B795	151	134	151	144	.	789	292	116	126	WILD
B819	158	135	151	146	152	790	289	111	125	WILD

우선 판별분석에 대한 이해를 위하여 **HUM, ULN 두 변수만 측정하였다고** 가정하고 판별분석을 실시하자. 판별분석에서는 각 개체의 측정 변수 중 하나라도 결측치가 있으면 그 개체는 분석에서 제외된다. 우선 판별분석의 이해를 돕기 위하여 산점도를 그려보자. 실제 판별분석에서는 산점도를 그릴 필요는 없다.

만약 아래와 같이 직선을 임의로 그어 판별식으로 사용하면 야생 세 마리가 사육으로 오분류 된다. 판별 변수가 2 개일 경우에는 가능하다.



Fisher 의 판별분석을 실시해 보자. 판별분석 옵션 창이 열려 집단 변수를 지정하려고 보니 TYPE 변수가 나타나지 않는다. 이는 문자열로 지정되었기 때문이다. 그래서 변환(T)➔코딩변경(R)을 통하여 WILD=1, DOMESTIC=2 로 변환하였다. “범위지정(D)” 설정

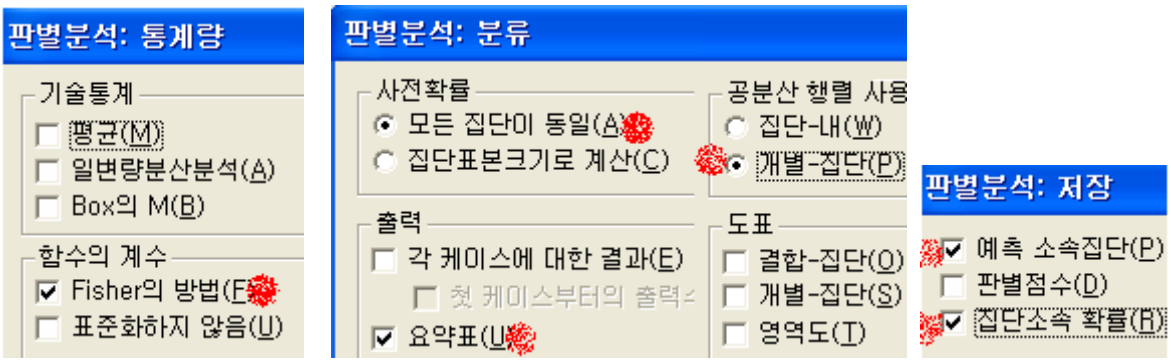


함수의 계수는 판별에 사용되는 식을 얻기 위한 것이다. 표준화는 원 변수를 표준화 했을 때 식의 계수이다. 일반적으로 분석이나 해석에 사용하지 않으므로 선택할 필요는 없다.

사전확률에서 표본이 모집단의 비율에 맞추어 층화추출(stratified sampling) 되었다면 “집단 표본크기로 계산” 옵션을 선택하면 된다. 공분산 행렬은 사용된 분류 변수의 단위가 비

슷하면 통합 분산을 이용하자(등분산 검정은 추후 논의). "요약표"는 오분류 표 출력하라는 옵션이다.

"집단소속 확률"은 개체가 집단에 소속될 확률을 데이터에 저장하는 옵션이다. 개체는 소속 확률이 가장 큰 집단에 소속되는데 이에 대한 정보는 "예측 소속집단"을 선택하면 데이터에 변수로 저장된다.



TYPE	group	Dis_1	Dis1_1	Dis2_1
WILD	1	1	.64586	.35414
WILD	1	1	.87073	.12927
WILD	1	.	.	.
WILD	1	1	.82165	.17835
WILD	1	1	.88689	.11311
WILD	1	.	.	.
WILD	1	2	.48784	.51216
WILD	1	1	.94226	.05774
WILD	1	1	.90192	.09808
WILD	1	1	.96288	.03712

사전 확률은 0.5 로 설정되었음을 알린다. GROUP=1 인 칠면조가 17 마리, 사육인 칠면조가 20 마리이다. 칠면조 판별 함수는 다음과 같이 쓸 수 있다.

$$Y_{야생} = -520.85 + 3.4 * HUM + 3.48 * ULN \ , \ Y_{사육} = -470.19 + 3.33 * HUM + 3.21 * ULN$$

집단에 대한 사전확률

group	사전확률	분석에 사용된 케이스	
		가중되지 않음	가중됨
1	,500	17	17,000
2	,500	20	20,000
합계	1,000	37	37,000

분류 함수 계수

	group	
	1	2
HUM	3,401	3,325
ULN	3,481	3,212
(상수)	-520,852	-470,190

Fisher의 선형 판별함수

위의 판별식에 의해 GROUP=1(야생)이 사육으로 잘못 분류된 칠면조는 3 마리(오분류 비율 17.6%), 사육인데 야생으로 분류한 칠면조는 5 마리(오분류 비율 26.3%)이다. 전체

오분류 비율은 21.6%이다. 어느 칠면조가 오분류 되었는지는 데이터의 집단 분류 결과를 보면 된다.

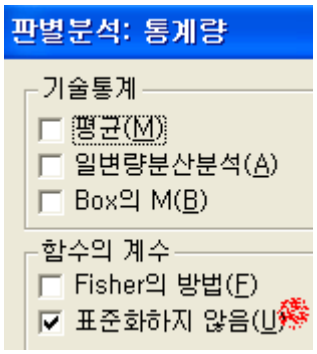
분류결과[a]

			예측 소속집단		
		group	1	2	전체
원래값	빈도	1	14	3	17
		2	5	15	20
	%	1	82.4	17.6	100.0
		2	25.0	75.0	100.0

a. 원래의 집단 케이스 중 78.4%이(가) 올바로 분류되었습니다.

집단 분류 결과에 대한 산점도를 그리면 다음과 같다. 페이지 123 의 산점도와 비교하면 어느 칠면조가 오분류 되었는지 알 수 있다. 그런데 이상하다. 앞에서 내가 임의로 그은 판별식에 비해 오분류 비율이 높다. 왜 그러지? 판별분석은 집단 평균 점으로부터의 거리 개념으로 개체를 분류하기 때문이다.

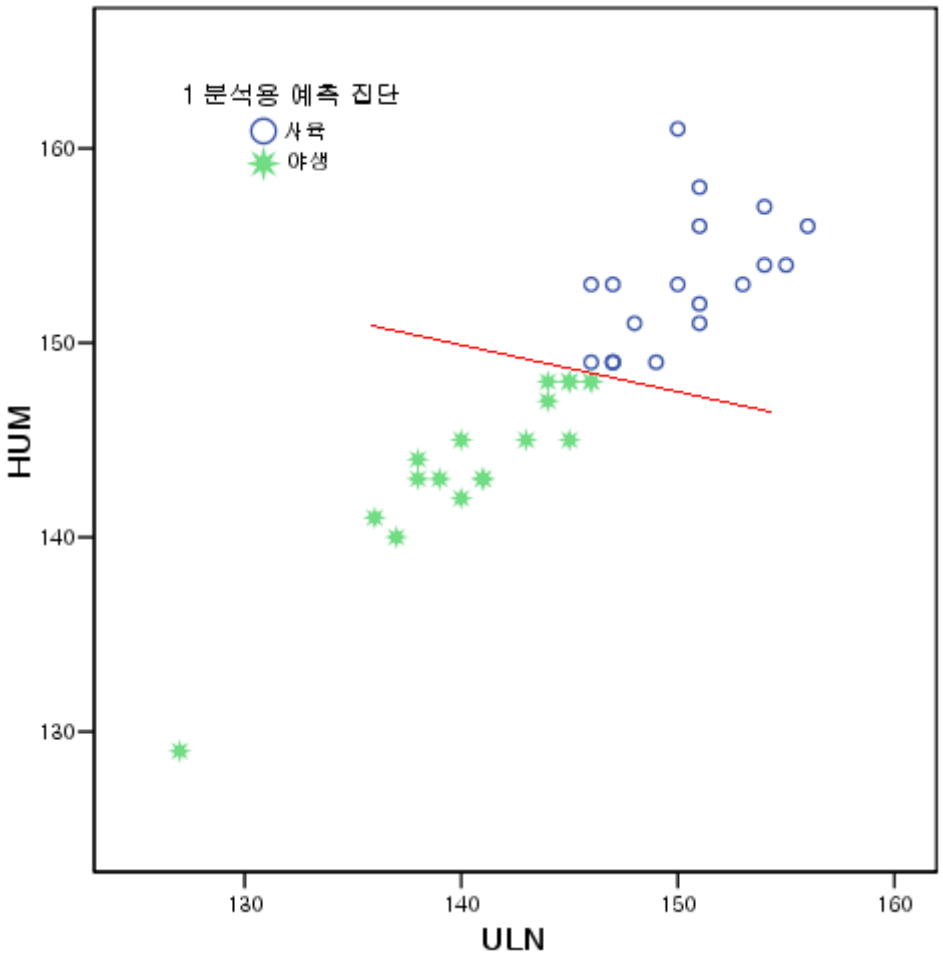
공분산 행렬을 사용할 때 통합 분산을 사용할 수 있는지 알아보는 공분산 동일성 검정을 하려면 다음 옵션만 설정해 주면 된다. 유의확률이 0.715 이므로 귀무가설이 채택되어 **등분산** 가정이 만족한다. 공분산을 통합 분산으로 사용할 수 있다. (이미 앞에서 통합 분산 사용)



검정 결과

Box의 M		.137
F	근사법	.133
	자유도1	1
	자유도2	3603.527
	유의확률	.715

정준 판별 함수의 모집단 공분산행렬이 동일하다는 영가설을 검정합니다.



4.2.2 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)

데이터 제일 아래 새로운 개체 2 개를 다음과 같이 입력한 후 판별분석을 실시하면 새로운 데이터는 판별식 구하는데 사용하지 않고 새로운 데이터가 각 집단에 속할 확률과 어느 집단에 속하는지 출력된다.

ID	HUM	RAD	ULN	FEMUR	TIN	CAR	D3P	COR	SCA	TYPE	group
L770	144	129	138	130	134	790	300	97	118	DOM	2
L774	143	128	141	130	137	800	300	98	123	DOM	2
새1	145	.	150	.	.	.	.	.	.		.
새2	150	.	145	.	.	.	.	.	.		.

분류 결과를 보면 기존의 37 마리 칠면조에 대한 분류는 전과 동일하다. 왜냐하면 판별분석에는 새로운 칠면조 2 마리 정보는 사용되지 않았다. 그래서 오분류 비율은 전과 동일

하게 78.4%이다. 새로운 칠면조 1 은 야생(야생 소속 확률=0.69), 칠면조 2 는 사육(사육 소속 확률=0.535)으로 분류되었다.

분류결과[a]

		group	예측 소속집단 1	예측 소속집단 2	전체
원래값	빈도	1	14	3	17
		2	5	15	20
		집단화되지 않은 케이스	1	1	2
	%	1	82.4	17.6	100.0
		2	25.0	75.0	100.0
		집단화되지 않은 케이스	50.0	50.0	100.0

a. 원래의 집단 케이스 중 78.4%이(가) 올바로 분류되었습니다.

ID	HUM	RAD	ULN	REM	TIN	CAR	DPR	COR	SCA	TYP	group	Dis_1	Dis1_1	Dis2_1
L770	144	*	*	*	*	*	*	*	*	D	2	2	.06099	.93901
L774	143	*	*	*	*	*	*	*	*	D	2	2	.13208	.86792
새1	145	.	*	.	.	.	.	.	.		.	1	.69465	.30535
새2	150	.	*	.	.	.	.	.	.		.	2	.46490	.53510

4.2.3 일반적인 경우

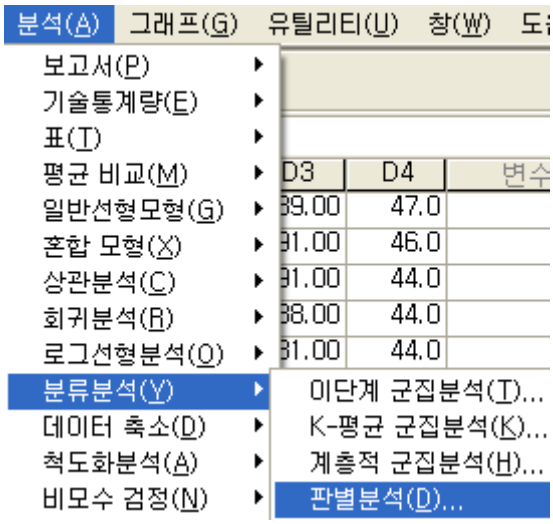
밀(Wheat) 종류에는 Arthur 종(soft 한 밀)과 Arkan 종(hard 밀)이 있고 Group 1, 2 과 Group 3, 4 는 서로 다른 지역이다. 그러므로 4 개의 집단이 존재한다. 밀에 대해 다음 항목의 길이를 조사하였다. n=172

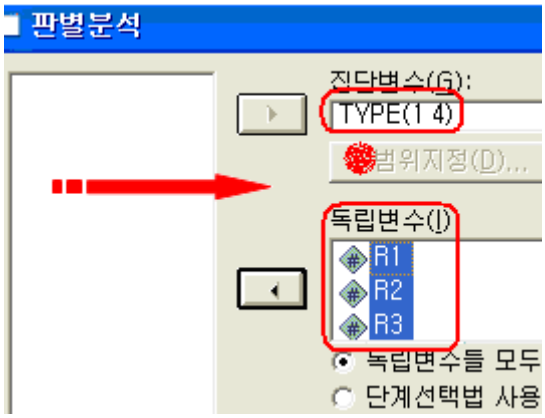
밀의 오른쪽(Right) 면에서 면적(Area R1), 원주(Perimeter R2), 길이(Length R3), 폭(breadth R4) 그리고 아래쪽(down)에서 면적(Area D1), 원주(Perimeter D2), 길이(Length D3), 폭(breadth D4)을 조사하였다. ■🖫WHEAT.SAV■

[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998]

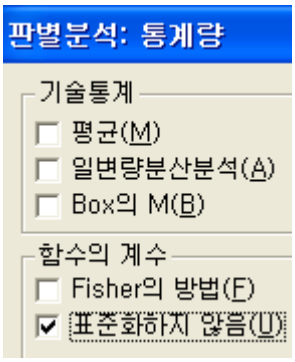
TYPE	R1	R2	R3	R4	D1	D2	D3	D4
1	54.45	219.0	89.00	43.0	56.60	226.0	89.00	47.0
1	55.15	221.0	91.00	46.0	56.26	224.0	91.00	46.0
1	53.92	223.0	90.00	44.0	55.09	223.0	91.00	44.0
1	52.23	212.0	87.00	41.0	53.54	215.0	88.00	44.0
1	51.56	207.0	78.00	42.0	52.98	211.0	81.00	44.0

집단 변수와 독립(판별) 변수를 지정해 준다. 집단 변수를 지정하면 “범위지정” 옵션이 설정 가능하게 된다. 여기서 분류하고자 하는 집단(범주)을 설정해 주면 된다.





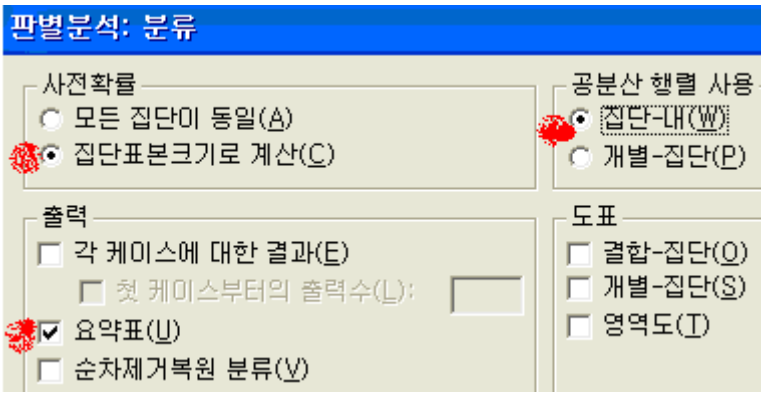
우선 등분산 가정을 검정해 보자. 귀무가설이 기각되어 등분산 가정이 무너지므로 집단-내 공분산을 사용한다.

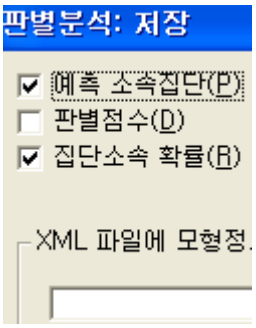


검정 결과

Box의 M		69.195
F	근사법	3.716
	자유도1	18
	자유도2	81222.713
	유의확률	.000

표본을 추출할 때 층화추출 했다고 가정하여 사전 확률 옵션을 “집단표본크기로 계산”을 선택하였다. 등분산 가정이 무너졌으므로 집단-내 공분산 사용하였다.





전체 오분류 비율은 69%였다. 특히 지금 우리가 구한 Fisher 판별식은 밀의 종류 중 1과 2에 속한 밀을 오분류할 가능성이 높다.

분류결과[a]

		TYPE	예측 소속집단 1	2	3	4	전체
원래값	빈도	1	23	1	12	0	36
		2	3	17	3	13	36
		3	8	0	39	3	50
		4	1	6	4	39	50
	%	1	63,9	2,8	33,3	,0	100,0
		2	8,3	47,2	8,3	36,1	100,0
		3	16,0	,0	78,0	6,0	100,0
		4	2,0	12,0	8,0	78,0	100,0

a. 원래의 집단 케이스 중 68.6%이(가) 올바로 분류되었습니다.

새로운 밀이 입고되었는데 생산지역을 모른다고 하자. 그 밀에 대해 8 개 항목을 측정한 후 관측치를 데이터 마지막 행에 넣는다. 최종 판별분석 과정을 실행하면 새로운 밀이 어느 지역 밀인지 판별하게 된다.

TYPE	R1	R2	R3	R4	D1	D2	D3	D4	Dis	Dis	Dis	Dis	Dis	Dis_2
4	55.34	231.0	97.00	43.0	53.06	230.0	98.00	41.0	4	0	0	0	1	4
4	59.65	230.0	89.00	53.0	56.25	227.0	98.00	46.0	4	0	0	0	1	4
.	50.00	200.0	90.00	50.0	50.00	250.0	100.0	50.0	.	.	.	.	.	2

4.3 단계별 판별분석

판별분석에서 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.

4.3.1 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.

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.

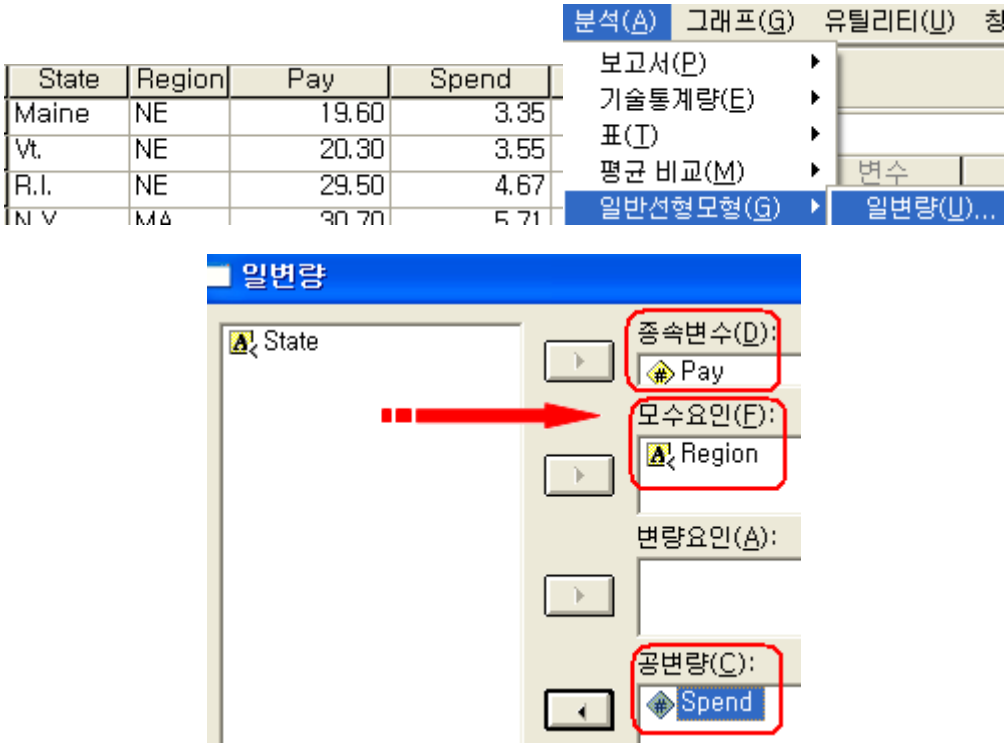
🕮 **공분산 분석:** 수학 강의에 대한 새로운 교육방법이 제안되었다. 기존의 교육 방법보다 나은지 알아보기 위하여 각 20 명씩 2 개의 그룹을 만들어 하나의 그룹에는 새로운 교육 방법, 다른 그룹에는 기존의 교육 방법을 적용해 보자. 그룹 학생들간에는 차이가 있을 것을 예상하여 교육 전 수학 시험을 보았다. 일전 기간 교육 후 수학 능력 시험을 봐 그 성적의 차이가 있는지 분석하였다. 교육 후 점수(Y)가 그룹(새 교육/기존 교육)간 차이가 있는지 알아보려면 분산분석(ANOVA). 그러나 교육 전 이들의 수학 능력이 고려되지 않았다. 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.

EXAMPLE 4-1 공분산 분석

http://lib.stat.cmu.edu/DASL/Datafiles 주 이름, 지역, 연봉, 교육 재정 데이터. ■ 🖫PAY.SAV■

지역별 선생들의 연봉 차이를 알아보기 위하여 수집한 데이터이다. 그러나 연봉은 주 교육 재정에 따라 차이가 있을 것으로 판단하여 이를 공변량으로 사용하였다. 우리의 관심은 주별 연봉 차이이다. 공변량은 관계 분석을 보다 올바르게 하기 위하여 고려되는 변수이다.

State	Region	Pay	Spend
Maine	NE	19.60	3.35
Vt.	NE	20.30	3.55
R.I.	NE	29.50	4.67
N.Y.	MA	30.70	5.71



지역별 연봉의 차이는 유의하지 않다. 유의확률은 0.104.

개체-간 효과 검정

종속변수: Pay

소스	제 III 유형 제곱합	자유도	평균제곱	F	유의확률
수정 모형	679,747[a]	9	75,527	14,900	,000
절편	368,885	1	368,885	72,775	,000
Spend	382,249	1	382,249	75,412	,000
Region	73,256	8	9,157	1,807	,104
오차	207,822	41	5,069		
합계	31485,620	51			
수정 합계	887,568	50			

a. R 제곱 = ,766 (수정된 R 제곱 = ,714)

4.3.2 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.

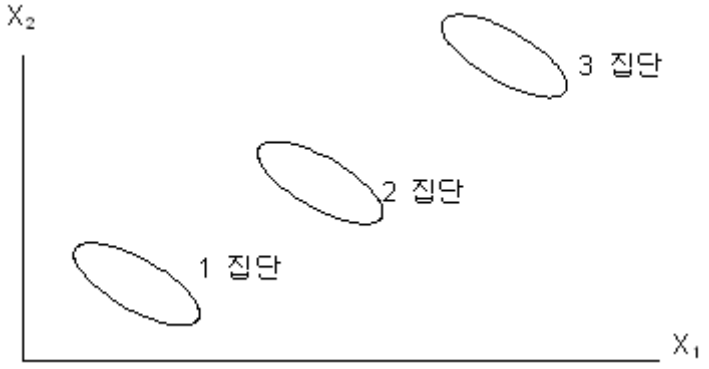
(2)같은 방법으로 변수를 하나씩 제거해 간다. 집단의 SS3 의 F-값이 모두 유의하면 (p-값이 유의수준보다 작으면) 제거를 멈춘다. 일반적으로 유의수준은 0.15 로 한다. SAS 에서는 SLS(Significant Level for Stay) 옵션을 설정할 수 있다.

4.3.3 Stepwise 방법

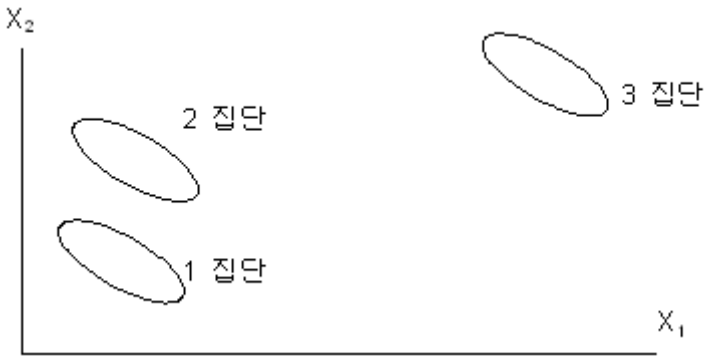
Forward 방법과 매우 유사하다. 일단 선택된 변수들도 다른 변수가 들어간 상태에서 유의성 검정을 하여 새로운 변수보다 덜 유의하면 제거된다. 즉 처음에는 가장 유의하였지만 여러 변수들이 선택된 상황에서는 유의 정도가 떨어질 수 있다. 변수 선택을 위하여 변수를 넣었다 뺐다 하는 반복이 많아 계산이 복잡하고 번거롭지만 컴퓨터 하드웨어, 소프트웨어 발달로 인하여 현재는 가장 많이 사용되는 방법이다.

4.3.4 변수 선택 시 주의점

변수의 수가 15 개 이상인 경우 Backward 방법, 15 개 미만이 경우는 Stepwise 방법을 사용하는 것을 권한다. 판별분석에서 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.



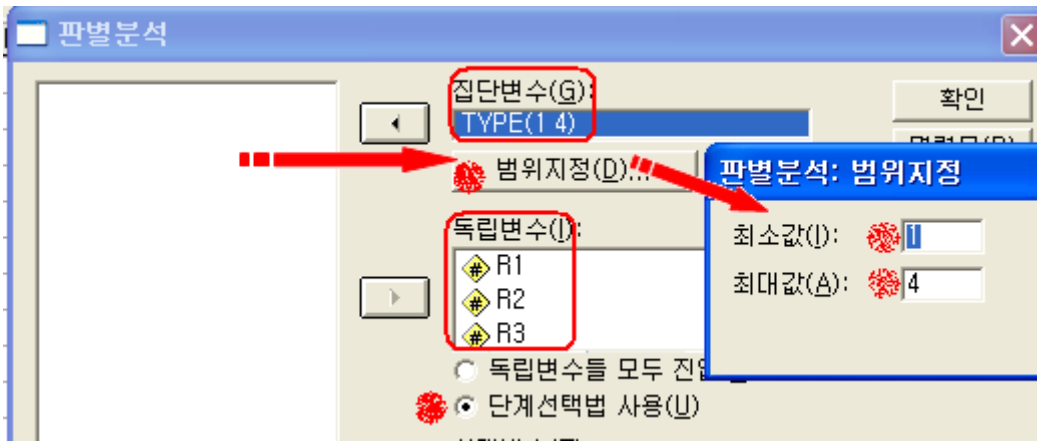
만약 산점도가 위와 같다면 실제로는 X_2 가 더 나은 판별을 하지만 변수 선택 방법으로 선택하면 X_1 이 선택될 것이다. 그러므로 변수 선택 방법에 의해 선택된 변수가 판별을 잘 한다는 보장은 없다. 그럴지라도 변수 선택 방법은 하나의 좋은 기준을 제시한다.

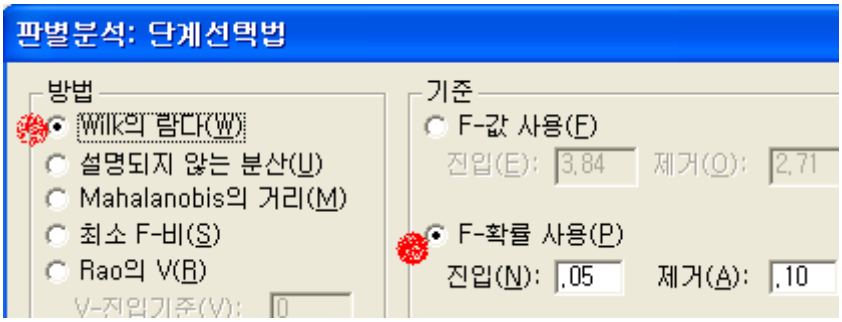


판별분석의 주요 목적은 개체들을 오분류 없이 분류하는 것이므로 다양한 변수 선택 방법을 사용해 보고 각각에 대해 (1)측정 변수 모두 사용하여 오분류 비율을 구하고 (2)변수 선택 후 오분류 비율을 계산하여 오분류가 가장 적은 판별분석 방법과 측정 변수 군을 이용하는 것이 올바른 접근 방법이다.

4.3.5 변수 선택 방법 예제

유의한 변수 삽입 유의확률은 0.05, 제거 유의확률은 0.1 로 하였다. 나머지 옵션은 이전과 동일하다.





개체를 판별하는데 유의한 변수들은 R1, R2, D1, D3. D4 이다. 8 개 변수를 모두 사용했을 때에 비해 오분류가 31%에서 34%로 조금 증가하였다. 그러나 새로운 개체를 분류하는데 5 개 변수만 필요하니 경제적이다.

판별에 유의한 변수만으로 개체를 분류하는 것이 오분류 비율 면에서도 더 효율적일 경우도 빈번히 발생하므로 두 방법 모두 사용하여 오분류 비율과 오분류 형태를 보고 분석자가 어느 판별 함수를 사용할지 판단하면 된다.

분류결과[a]

		TYPE	예측 소속집단 1	2	3	4	전체
원래값	빈도	1	27	1	8	0	36
		2	4	20	2	10	36
		3	12	0	36	2	50
		4	1	17	2	30	50
	%	1	75.0	2.8	22.2	.0	100.0
		2	11.1	55.6	5.6	27.8	100.0
		3	24.0	.0	72.0	4.0	100.0
		4	2.0	34.0	4.0	60.0	100.0

R1
R2
D1
D3
D4

a. 원래의 집단 케이스 중 65.7%이(가) 올바로 분류되었습니다.

4.4 판별분석 2

4.4.1 로지스틱 판별분석

판별분석은 판별 변수가 모두 측정형인 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 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 있다. SPSS 는 이진형 로지스틱 회귀분석만 제공한다.

▣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_i 가 이진형인 경우(자료가 0, 1 만 존재) OLS 에 의한 계수 추정은 무의미 하다.

ODDS 변환: $odd = \frac{p}{1-p}$ **:** 어떤 사건이 발생할 가능성[p=0.5일 경우 1이다. 기준]으로 해석 될 수 있다. 한국이 2002년 16강에 들어갈 확률 0.1이면 1/9이 Odds이다. ▶ 1$ betting 에서 이기면 9$ return 브라질이 2002년 16강에 들어갈 확률 0.8이면 4가 Odds이다. ▶ 4$ betting에서 이기면 1$ return

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 변환 $p_i^* = \frac{p_i}{1-p_i}$ 이다.

확률 p_i 가 (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 로지스틱 모형이라 한다.

$$y_i = \ln(\frac{p_i}{1-p_i}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + .. + \beta_p x_{pi} + e_i$$

위의 모형을 다시 쓰면 다음과 같다.

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

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. 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.

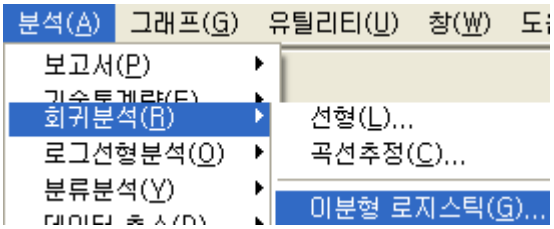
▣분석하기

EXAMPLE 4-2 로지스틱 회귀분석

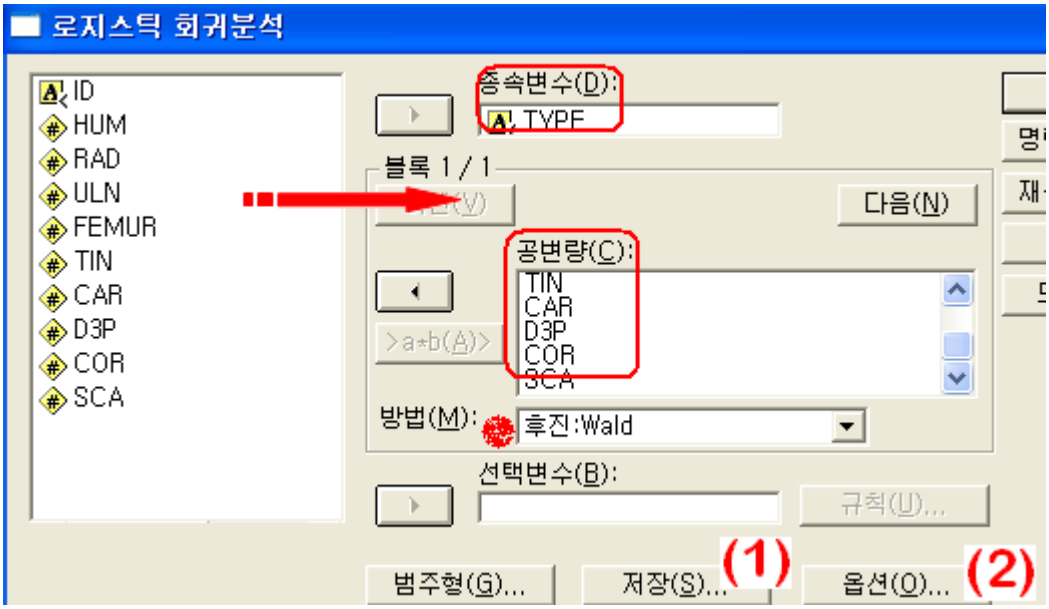
미국 Kansas 주립대학 Dr. Michael Finnegan 교수는 야생 칠면조와 사육 칠면조를 구별하기 위하여 수컷 칠면조 82 마리에 대해 9 개 항목에 대한 측정치를 조사하였다.

[Applied Multivariate Methods for Data Analysts, Dallas E. Johnson, 1998] ■ TURKEY.SAV■

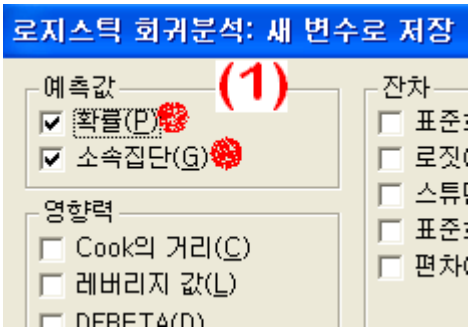
메뉴에서 로지스틱 회귀분석에서 종속변수와 설명변수를 설정한다. “1”의 범주형 옵션은 설명변수 중 범주형 변수를 지정한다. 로지스틱 회귀분석에서 범주형 설명변수는 회귀분석과 마찬가지로 가변수(지시변수)처럼 다루어야 한다.

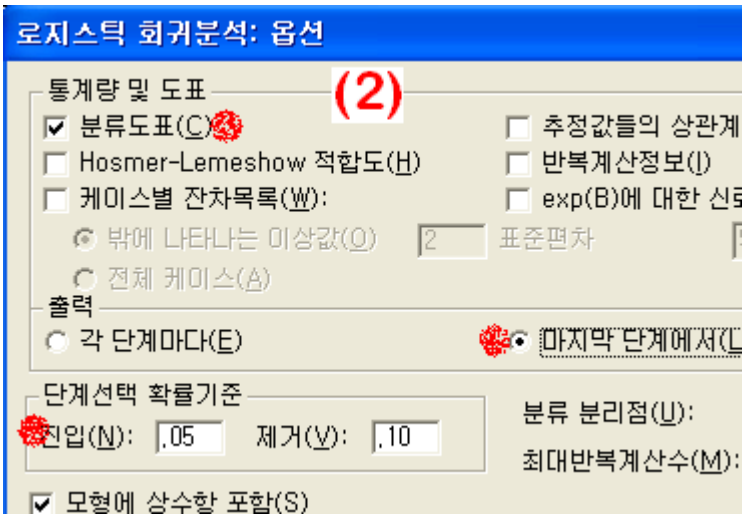


변수 선택방법으로는 후진 제거(WALD 통계량 이용) 방법을 사용하였다.



"2"에서는 성공 확률과 어느 집단으로 판별되는지 데이터에 변수로 저장되도록 설정하였다. "3"에서는 오분류 표와 추정 결과 최종 단계만 출력되게 하였다.





출력 결과가 황당하게 많다. 필요한 부분만 설명하도록 하겠다. “종속변수 코딩” 결과는 사육이 성공으로 사용하였다는 말이다. 데이터에 저장된 성공 확률은 사육일 확률이다. 3 번째 칠면조의 성공(사육) 확률이 0.98 이므로 사육으로 판별된다.

종속변수 코딩

원래 값	내부 값
WILD	0
DOMESTIC	1

TIN	CAR	D3P	COR	SCA	TYPE	PRE_1	PGR_1
151	817	305	102	128	WILD	.00004	WILD
155	814	305	111	137	WILD	.00000	WILD
145	775	.	106	128	WILD	.98917	DOMESTIC
.	789	292	116	126	WILD	.	
152	790	289	111	125	WILD	.00034	WILD
149	789	.	111	123	WILD	.	
147	767	287	106	123	WILD	.30392	WILD

변수 선택 과정이 출력되었다. 왜 모두 다 출력하는지... 9 단계가 최적 단계이다. 10 단계 결과를 보면 새로 들어간 FEMUR 변수는 유의하지 않다. B 는 표준화 회귀 계수를 의미한다. TIN 의 부호가 음이므로 TIN 의 크기가 클수록 성공 확률이 높아진다(사육 칠면조일 가능성이 높다). 오분류 비율은 13%로 Fisher 의 판별분석 방법 31%에 비해 줄었다.

그런데 하나만 사용하는 것이 다소 꺼림직하지 않능가? 그런 경우... 유의수준을 0.2 로 올려보자. 7 단계(다음 페이지) 결과를 보면 된다. ULN, TIN, COR 이 칠면조 판별에 유의하다. TIN, COR 은 작을수록, ULN 은 클수록 사육일 가능성이 높다. 표준화 회귀계수를 보면 ULN, TIN 의 영향이 사육이 가능성에 미치는 영향력이 COR 의 두 배이다. 세 변수를 판별 변수로 사용하였을 경우 정분류 비율이 93.5%이다. 이 방법에 의해 판별할 경우 오분류 비율이 가장 낮으므로 칠면조를 분류하는데 최적 판별방법이 된다.

방정식에 포함된 변수[c]

		B	S.E.	Wald	자유도	유의확률	Exp(B)
1 단계	HUM	-.757	18954.895	.000	1	1.000	.469
	RAD	-9.196	19379.886	.000	1	1.000	.000
	ULN	6.402	3518.262	.000	1	.999	603.035
(중간생략)							
7 단계	TIN	-1.056	.617	2.935	1	.087	.348
	상수	82.992	46.995	3.119	1	.077	1.104E+36
9 단계	TIN	-.477	.179	7.064	1	.008	.621
	상수	69.404	26.129	7.055	1	.008	1.386E+30
10 단계	FEMUR	1.379	1.029	1.797	1	.180	3.971
	TIN	-3.361	2.457	1.872	1	.171	.035
	상수	301.566	222.224	1.842	1	.175	9.298+130

분류표[a]

관측			예측값 TYPE WILD	예측값 TYPE DOMESTIC	분류정확 %
1 단계	TYPE	WILD	14	0	100,0
		DOMESTIC	0	17	100,0
	전체 %				100,0
2 단계	TYPE	WILD	14	0	100,0
		DOMESTIC	0	17	100,0
	전체 %				87,1
9 단계	TYPE	WILD	12	2	85,7
		DOMESTIC	2	15	88,2
	전체 %				87,1
10 단계	TYPE	WILD	13	1	92,9
		DOMESTIC	0	17	100,0
	전체 %				96,8

(중간생략)

7 단계	ULN	1,157	,758	2,327	1	,127	3,180
	TIN	-1,180	,758	2,420	1	,120	,307
	COR	-,533	,404	1,738	1	,187	,587
	상수	58,890	47,648	1,528	1	,216	3,765E+25

7 단계	TYPE	WILD	13	1	92,9
		DOMESTIC	1	16	94,1
	전체 %				93,5

▣새로운 개체 판별분석

새로운 개체 판별은 이전과 동일하다. 판별 변수를 ULN, TIN, COR 로 하고 새로운 개체에 대한 데이터를 마지막 행에 입력한 후 최종 판별분석을 실시하면 집단이 판별된다. 물론 집단 변수는 결측치로 입력한다.

4.4.2 다른 판별분석 방법

▣정준 판별분석

Fisher 에 의해 제안된 방법으로 Fisher's between-within method 라고 불리는 방법이다. 판별 변수들의 유용한 정보를 모두를 포함한 정준 (Canonical) 변수를 이용하여 판별분석을 실시한다. 판별 변수들의 수가(p) 너무 많아 판별 결과에 대한 해석이 곤란한 경우 p -차원 공간에서의 개체들의 집단 평균들을 저 차원 공간으로 변환시켜 처리하는 판별분석 방법이다. 개체 분류가 목적이 아니라 개체 분류 해석을 위해 저 차원(BOX-PLOT 이나 산점도)으로 표현하는데 있으므로 엄밀히 말하면 판별분석은 아니다. 새로운 변수

(정준 변수)에 대한 해석이 가능하든 아니든 집단들 사이의 실제 거리를 저차원으로 축소하여 시각화 할 수 있다는 장점이 있다. 차원을 줄인다는 의미에서 보면 주성분분석과 유사해 보이지만 계산 방법은 전혀 다르다.

▣제일 정준 함수

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

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

$\max_{\underline{b} \neq 0} \frac{\underline{b}'B\underline{b}}{\underline{b}'(B+W)\underline{b}}$ 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 평균과 의 거리는 $d_i = |\underline{b}_1'\underline{x} - \underline{b}_1'\underline{\hat{\mu}}_i|$, $i = 1,2,...,m$ 이다.

▣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 평균과 의 거리는 다음과 같이 계산된다. $d_i = (\underline{b}_1'\underline{x} - \underline{b}_1'\underline{\hat{\mu}}_i)^2 + (\underline{b}_2'\underline{x} - \underline{b}_2'\underline{\hat{\mu}}_i)^2$, $i = 1,2,...,m$

▣차수 결정

주성분분석과 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.

▣K Nearest Neighbor 판별분석

모집단이 정규분포를 따르지 않는 경우 사용하는 비모수 판별분석 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.

SPSS 는 이 방법을 제공하지 않고 있다.

▣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. SPSS 에는 ANSWER tree TOOL 에 속해 있다.

EXERCISE

[데이터 출처: http:// lib.stat.cmu.edu/DASL] ■🖫CAR.SAV■

자동차의 (US, non-US) 생산국을 집단으로 하여 Fisher 의 판별분석(변수 선택/모든 선택), 로지스틱 판별분석 방법 중 오분류가 가장 적은 방법을 고르시오.