

DA Concept (페이지 166)

Individual Directed Technique

- 측정 변수(항목)에 의한 개체 분류
 - 분류되어 있는 집단간의 차이를 의미 있게 설명해 줄 수 있는 독립변수들을 찾아내어
 - 변수의 선형결합으로 판별식(Discriminant function)을 만들어 낸다.
 - 이 판별식을 이용하여 분류하고자 하는 개체의 집단을 판별

데이터 유형

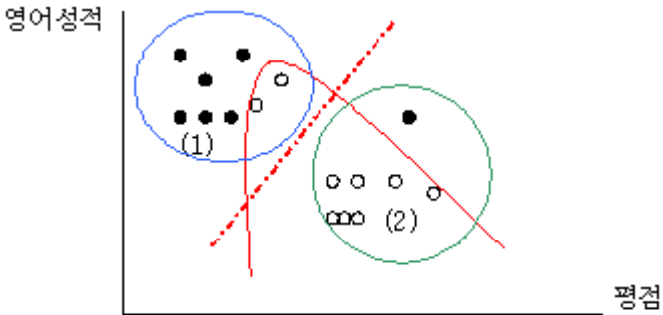
- 집단변수: 범주형 혹은 이진형
- 판별 변수: 측정형(등간 척도 포함)

사례

- SKT/KT/LGT 가입 고객 판별 변수 및 판별함수 유도
- 서비스 이용 불만고객 성향 분석

주성분 점수나 요인점수 이용 개체 판별?

- 원 변수를 축약한 주성분(요인) 점수의 산점도를 이용한 개체 집단 표현, 실제로는 판별 아님



Variable Directed Techniques

- 변수 축약: 주성분 점수
- 유사 변수그룹: 요인분석, 요인점수

• 개체분류
 ◻ 군집분석
 ◻ 판별분석

이름	취업 여부	어학 능력	학점	봉사활동
Kim	X	550	3.5	12 months
Lee	X	600	3.2	6 m
Park	X	700	4.0	0 m
Hong	O	850	3.8	24 m

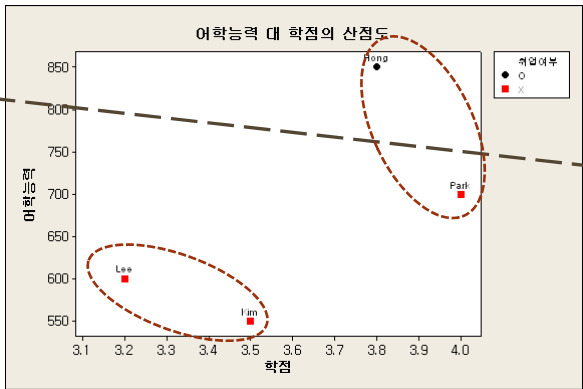
http://wolfpack.hnu.ac.kr



유사 분석

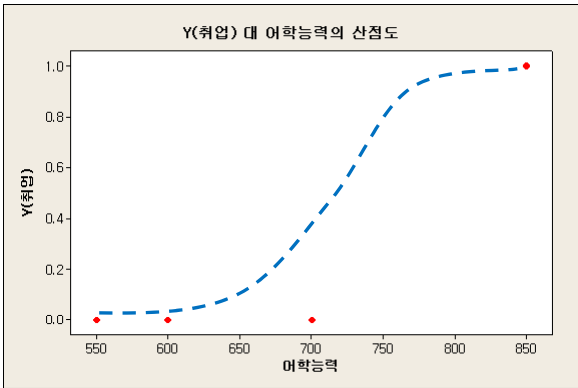
▪ 군집분석 clustering analysis

- (유사) 개체를 분류 (grouping)
- (상이) 데이터에는 집단을 구별하는 변수 없음 > 개체의 유사성(similarity)에 의해 개체 분류



▪ 회귀분석 Regression analysis

- (유사)
 - 집단 변수가 이진형 혹은 순서형 > 종속변수, Logistic Regression
 - 판별 변수와 회귀분석 독립변수 집단 차이 설명
- (상이)
 - 판별분석은 집단이 범주형인 경우에도 가능
 - 집단을 구별하는 판별식 유도(집단 분류), 회귀분석은 연결함수 이용 선형모형화(집단 소속 예측 확률)



http://wolfpack.hnu.ac.kr

Chapter 6. Discriminant Analysis



판별규칙 discriminant rule (페이지 169)

http://wolfpack.hnu.ac.kr

판별함수 (discriminant function)

- $R=f(X_1, X_2, \dots, X_p)$: 개체의 집단을 판별하는데 사용되는 판별변수의 함수
- 판별함수 집단이 2개(k=1집단, 2집단) 인 경우, 판별변수 X_1, X_2, \dots, X_p, Z : 판별점수, a_i 는 판별계수

$$Z = a_1X_1 + a_2X_2 + \dots + a_pX_p$$

판별함수 찾기

- 집단 내 분산에 비해 집단간 분산의 차이를 최대화하는 독립변수의 함수를 찾는다.

판별함수 개수

- $\text{Min}(\text{집단 개수}-1, \text{판별변수 개수})$

데이터 크기

- 관측치(개체)의 개수(데이터의 크기, 표본 크기)가 판별변수 개수의 20배 이상, 집단의 각 범주에 최소한 20개 관측치
- 위의 조건을 충족시키지 못하면 분석결과는 불안정(판별식을 구성하는 각 독립변수와 전체 판별식의 설명력과 예측력을 신뢰할 수 없다는 의미)해 짐

판별규칙

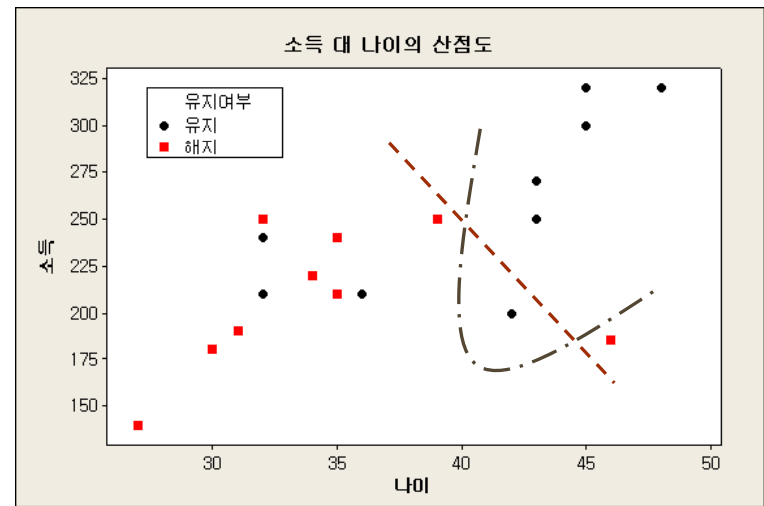
- 선형판별식: 두 집단의 분산이 같다는 가정

$$\underline{b}'x_0 - k > 0 \quad \underline{b}' = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} \quad k = (1/2)(\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

- Mahalanobis 거리: 두 집단의 분산이 같다는 가정

$$d_i = (x_0 - \underline{\mu}_i)' \Sigma^{-1} (x_0 - \underline{\mu}_i)$$

- 이차함수: 집단의 이분산 가정, 선형에 비해 경계선에 대한 유연성
- 우도함수: 판별변수의 분포가 정규분포 가정을 만족할 때
- SPSS에는 선형 판별식(등분산 옵션 선택)만 있음



오분류 misclassification (페이지 167, 171)

오분류

- 판별함수 신뢰 정도 평가하는데 사용

오분류율 (misclassification ratio)

- (오분류 개체 수) / (전체 개체 수) * 100
- 정분류율 (=1-오분류율): 회귀분석의 결정계수 R^2 개념

원 집단 \ 분류집단	집단1	집단2
집단1	정분류	오분류
집단2	오분류	정분류

오분류 계산 방법

Re-substitution 규칙

- 모든 개체 사용하여 판별식을 구하고, 이를 이용하여 오분류 비율 계산
- 간편하나 정분류율이 과대 추정 가능

Cross-validation 방법

- 개체 제외하고 판별식을 구하여 제외한 개체의 집단을 분류한다. 이 작업을 반복한다.
- 가장 많이 사용

테스트 데이터 이용

- 데이터를 이분하여, 한 데이터는 판별식(60~70%) 추정, 다른 데이터(40~30%)는 오분류율 계산에 사용
- 가장 정확한 오분류 계산, 어느 정도 대용량 데이터 확보 필요 (data mining에서)



비용함수 & 사전확률 (페이지 180)

http://wolfpack.hnu.ac.kr

■ 비용함수

- 오분류에 의한 비용함수 고려하여 판별식 선택
- 비용함수 선택
 - Equal Cost function (균등비용함수)
 - Ratio cost function (비례비용함수)
- 비용함수 고려 모형 복잡하므로 ECF 사용하여 오분류 표를 얻은 후 비용을 사후적 고려하는 것이 편리

■ SPSS에는 비용함수 설정 옵션 없음

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

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

■ 예제

■ 환자 마취 여부 판별: 판별식 1 사용이 적절

판별식1 ▷	마취 가능	마취 위험
마취 가능	95	10
마취 위험	5	90

■ 사전확률 prior

- 모집단의 구성비에 대한 정보가 있다면 이를 판별함수에 고려하는 것이 적절 (페이지 181)
 - priors EQUAL; 모집단의 각 집단 개체 비율이 일정
 - priors PROPORTIONAL; 표본의 집단 비율 사용
 - priors 설정; 모집단의 각 집단 개체 비율을 알고 있을 때
- 모집단의 개체 비율을 알고 있을 때는 설정하자.
- 모집단 집단 비율에 대한 확신이 없다면 표본 비율 사용

■ 설정?

- SAS, R : possible
- SPSS : not available

판별식2 ▷	마취 가능	마취 위험
마취 가능	90	5
마취 위험	10	90

Chapter 6. Discriminant Analysis



판별변수 선택 (페이지 191)

■ 개념

- 판별을 위해 선택된 변수가 판별 능력이 있나?
- (logic) 집단을 잘 분류한다? 집단 간 판별변수의 평균 차이 크다.
- (예제) (학점, 어학능력, 어학 연수기간)에 따른 취업집단 판별

■ 이유 parsimony 규칙

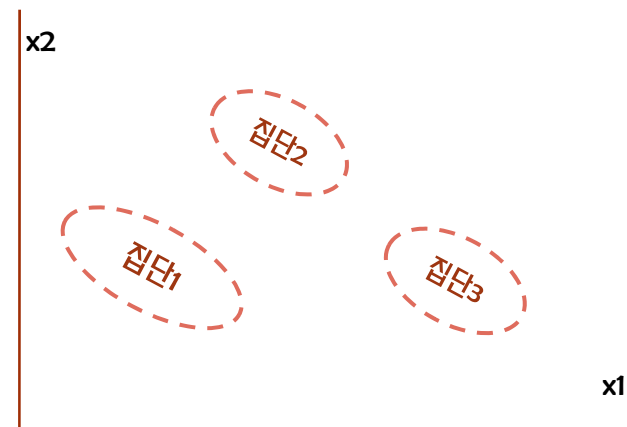
- 측정 오류 발생 가능성이 적고
- 새로운 개체 판별을 위해 측정해야 하는 변수 수가 적어 효율적.

■ 필요 개념 분산분석 및 공분산분석 개념

- 분산분석: 집단을 요인(처리효과)로 하고 판별변수를 종속변수로 하여 분산분석 결과 F값이 가장 큰 모형의 판별변수가 판별 능력이 가장 높은 판별변수
- 공변량 분석 (페이지 192): 분산분석 모형에서 측정형 설명변수를 공변량(covariate)라 한다. 종속변수의 사전점수나 이에 영향을 미칠 것으로 판단되는 측정형 변수, 실제 공변량 변수의 유의성에는 관심이 없고 요인의 유의성을 정확하게 검정하기 위하여 고려된다. (공변량=주 정부 지출, 지역적 고사 연봉의 차이 분석)

■ 판별변수 선택 방법

- Forward 방법
 - 판별 능력이 가장 유의한 변수 하나씩 입력
- Backward 방법
 - 판별 능력이 가장 유의하지 않는 판별 변수 하나씩 제거
- Stepwise 방법
 - Forward 방법과 유의하나 이미 삽입된 판별변수의 판별 유의성을 새로 입력된 변수들에 의해 검정
- 어느 변수가 판별 능력이 큰가? (페이지 195)



Fisher (선형) 판별분석 예제 (페이지 171)

http://wolfpack.hnu.ac.kr

▪ 데이터 TUKEY.txt

- 집단: wild, domestic
- 판별변수: 9개 부위 길이

```
> tk=read.table("turkey.txt",header=T)
> tk
      ID hum rad uln femur tin car d30 cor sca type
1   K766 . . . . . . . . . . WILD
2  N399 153 138 153 139 162 810 307 . . WILD
3  NEX1 . . . . . . . . . . WILD
4  NEX2 . . . . . . . . . . WILD
```

▪ Fisher 선형 판별분석 in R

```
> tk=read.table("turkey0.txt",header=T)
> library(MASS)
> da=lda(type~hum+rad+uln+femur+tin+car,data=tk,CV=T)
> names(da)
[1] "class" "posterior" "terms" "call" "
```

- class: 판별규칙에 의해 나누어진 사후 집단
- posterior: 사후 확률
- 왜 판별변수 모두 사용하지 않았나? 변수 수에 비해 개체 수가(9개 항목 모두 있는 개체 수 33 turkey0.txt) 너무 적어 collinearity 문제가 발생한다. 그래서 일부 변수만 사용하였다.
- CV=T: 사후확률 계산 옵션

▪ 판별분석 결과 object 내용

```
> da0=data.frame(tk$type,
+ da$class,da$posterior)
> da0
      tk.type da.class  DOMESTIC  WILD
1     WILD DOMESTIC 0.8097682520 1.902317e-01
2     WILD  WILD 0.0123787206 9.876213e-01
3     WILD  WILD 0.0155249415 9.844751e-01
4     WILD  WILD 0.2559034197 7.440966e-01
5     WILD  WILD 0.0007332552 0.002667e-01
```

▪ 오분류표 작성

```
> ct=table(da0$tk.type,da0$da.class)
> ct
```

	DOMESTIC	WILD
DOMESTIC	18	1
WILD	3	11

```
> diag(prop.table(ct,1))
      DOMESTIC  WILD
DOMESTIC 0.9473684 0.7857143
WILD     0.0526316 0.2142857
[1] 0.8787879
```

- 정분류 비율 87.8% (cross-validation 방법)
- 페이지 176 (두 개 변수, 0.78%)

▪ R에는 판별변수 선택 방법이 없다.

Chapter 6. Discriminant Analysis



Fisher 판별분석 예제 2 (페이지 17)

http://wolfpack.hnu.ac.kr

사전확률 설정

```
> da=lda(type~hum+rad+uln+femur+tin+car, data=tk, CV=T,
+ prior=c(0.4,0.6))
```

D=0.4, W=0.6 설정

```
> ct
          DOMESTIC WILD
DOMESTIC      16    3
WILD           2   12
> diag(prop.table(ct, 1))
DOMESTIC      WILD
0.8421053 0.8571429
> sum(diag(prop.table(ct)))
[1] 0.8484848
```

그래서 Wild로 분류된 개체 증가; 모집단에 대한 사전 정보가 있다면 사전 확률 지정이 유리

이차 판별함수

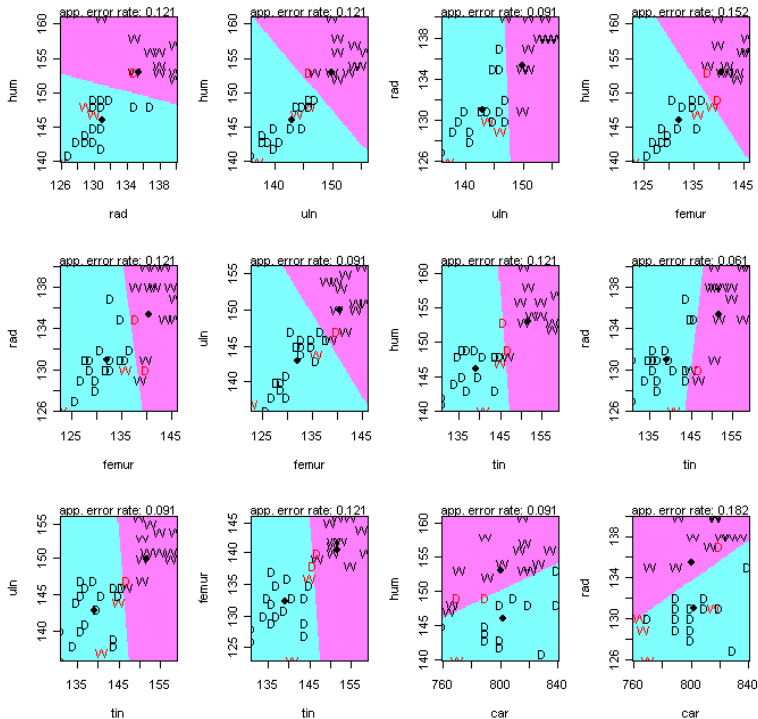
집단의 등분산 가정이 만족하지 않을 때 주로 사용

```
> da2=qda(type~hum+rad+uln+femur+tin+car,
+ data=tk, CV=T, prior=c(0.4,0.6))
> da3=data.frame(tk$type,
+ da$class, da2$posterior)
> ct=table(da3$tk.type, da0$da.class)
> ct
          DOMESTIC WILD
DOMESTIC      16    3
WILD           2   12
> diag(prop.table(ct, 1))
DOMESTIC      WILD
0.8421053 0.8571429
> sum(diag(prop.table(ct)))
[1] 0.8484848
```

판별분석 결과 표현

판별변수 산점도, 집단 id별

```
> library(klaR)
> partimat(type~hum+rad+uln+femur+tin+car,
+ data=tk, method="lda")
```



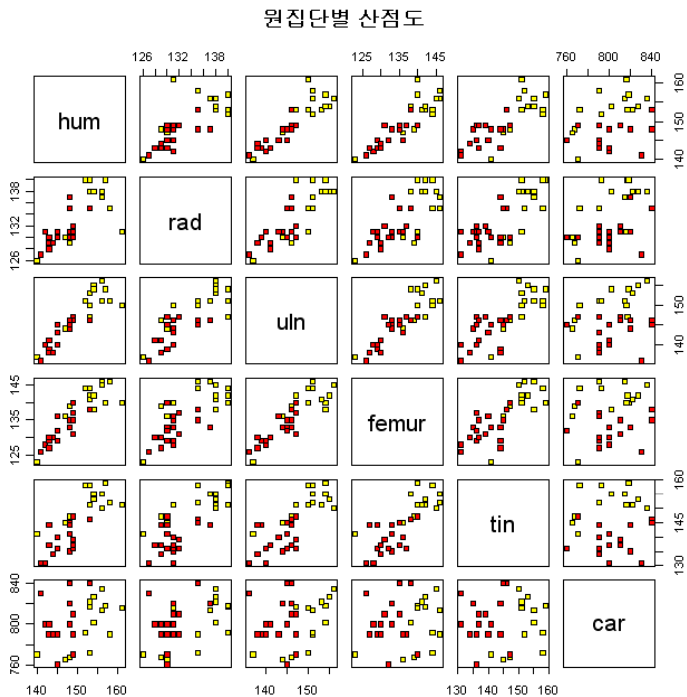
Fisher 판별분석 예제 3 (페이지 17)

산점도2

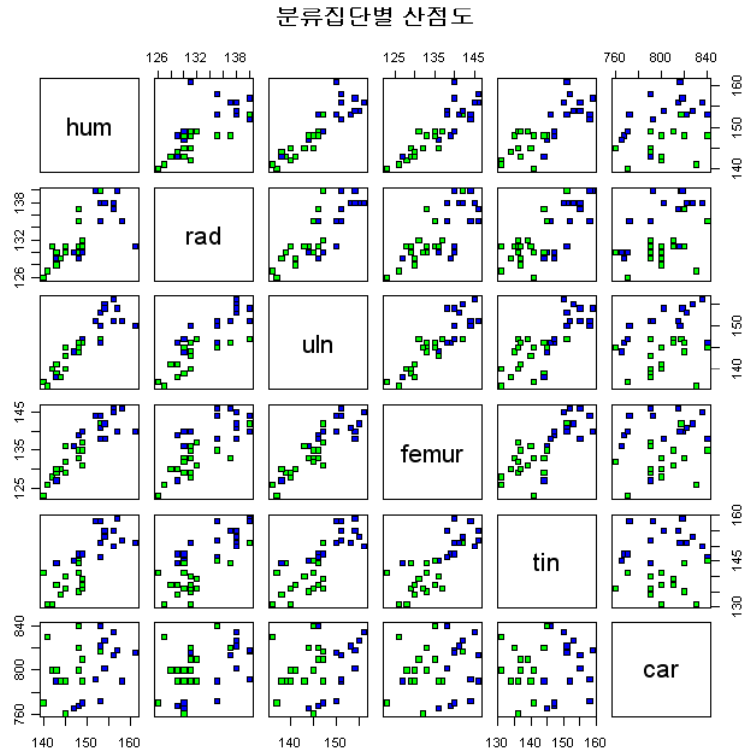
```

> #SCATTER PLOT
> pairs(tk[c("hum", "rad", "uln")],main="원집단별 산점도",
+ pch=22,bg=c("red", "yellow"))
+ [unclass(tk$type)]
> pairs(tk[c("hum", "rad", "uln")],main="분류집단별 산점도",
+ pch=22,bg=c("green", "blue"))
+ [unclass(da$class)]

```



분류집단 산점도



- 새로운 개체 분류 (페이지 199)
 - R에서 어떻게 할까?

http://wolfpack.hnu.ac.kr



판별식 활용 및 평가 (페이지 2)

오분류 개체 표현

- 판별변수별 산점도 (이전 슬라이드 참고)

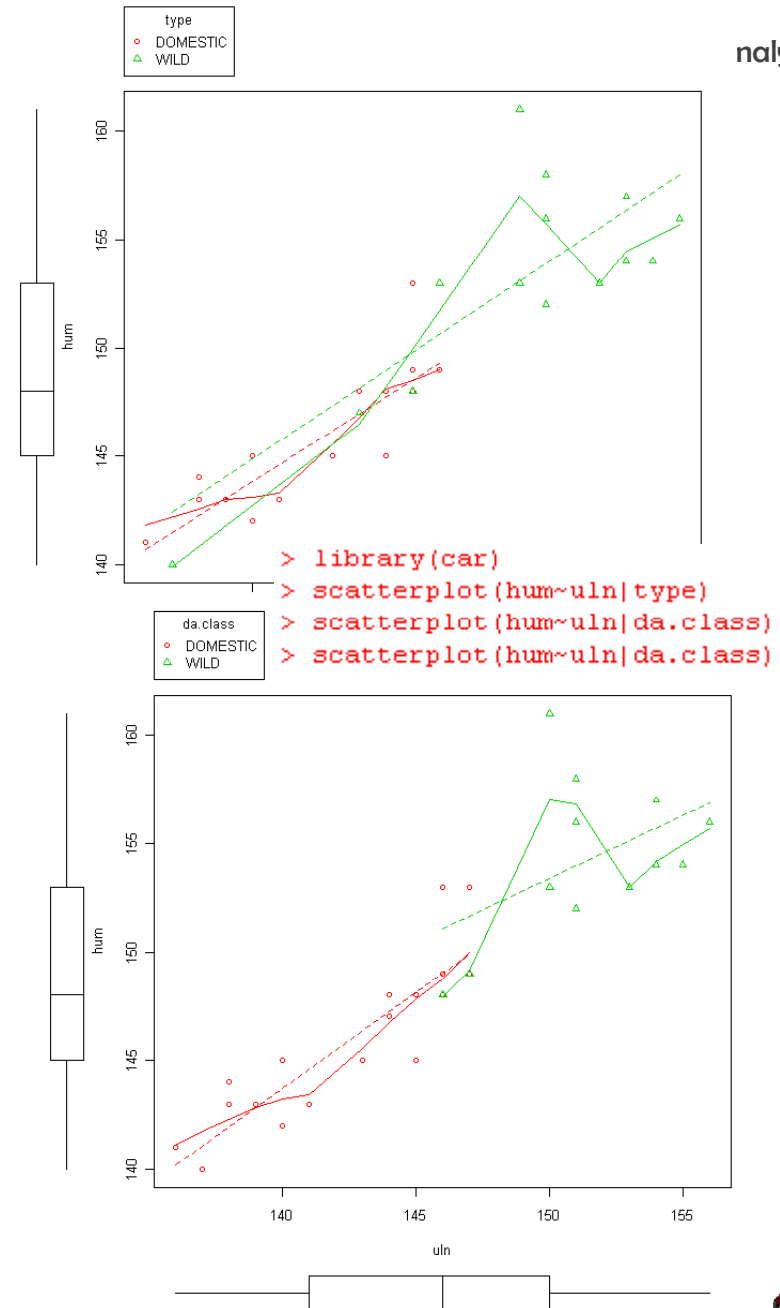
```
> da0$co[da.class=="Y"]="Y"
> da0$co[da.class!="Y"]="N"
> tk=read.table("turkey0.txt",header=T)
> library(MASS)
> da=lda(type~hum+rad+uln+femur+tin+car,data=tk,CV=T)
> da0=data.frame(cbind(tk,da$class,da$posterior))
> da0
```

	ID	hum	rad	uln	femur	tin	car	d3p	cor	sca	type
1	B710	153	140	147	142	151	817	305	102	128	WILD
2	B790	156	137	151	146	155	814	305	111	137	WILD

평균 이용

```
> da0$co[da.class=="type"]="Y"
> da0$co[da.class!="type"]="N"
>
> library(doBy)
> summaryBy(hum+uln+femur~co, data = da0,
+ FUN = function(x){c(m=mean(x))})
  co  hum.m  uln.m  femur.m
1  N 147.2500 143.7500 135.2500
2  Y 149.2414 146.1379 135.7586
```

```
> summaryBy(hum+uln+femur~type+da.class, data = da0,
+ FUN = function(x){c(m=mean(x))})
  type da.class  hum.m  uln.m  femur.m
1 DOMESTIC DOMESTIC 145.8889 142.6111 131.7778
2 DOMESTIC WILD 149.0000 147.0000 140.0000
3 WILD DOMESTIC 146.6667 142.6667 133.6667
4 WILD WILD 154.7273 151.9091 142.2727
```



nalysis

Chapter 6. Discriminant Analysis

http://wolfpack.hnu.ac.kr

Canonical DA 정준 판별분석 (페이지 200)

http://wolfpack.hnu.ac.kr

- Fisher 제안 Fisher's Between Within 방법
- 정준변수 활용
 - 원 판별변수의 선형 결합
 - 개체 집단차 차이를 최대화 하는 계수 구함
 - 제1정준변수: 개체 집단간 거리 최대화
 - 제2정준변수: 제1정준변수 계수와 직교, 개체간 거리 최대화
 - 정준변수간 상관계수는 0이다. $\frac{a' Ba}{a'(B+W)a}$
- 정준변수를 이용한 개체 저차원 표현
 - 차수 결정은 주성분 변수 개수 선택과 동일
 - 그러나 저차원에 표현하는 것이 목적이므로 2~3차원
- 예제 202 참고

- 다음 사이트 참고
 - http://bm2.genes.nig.ac.jp/RGM2/R_current/library/candisc/man/candisc.html

• fisher's iris 데이터



```
> iris=read.table("iris.txt",header=T)
> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1           50           33           14            2   setosa
2           64           28           56           22  virginica
3           65           28           46           15  versicolor
4           67           31           56           24  virginica
5           63           28           51           15  virginica
```

Chapter 6. Discriminant Analysis



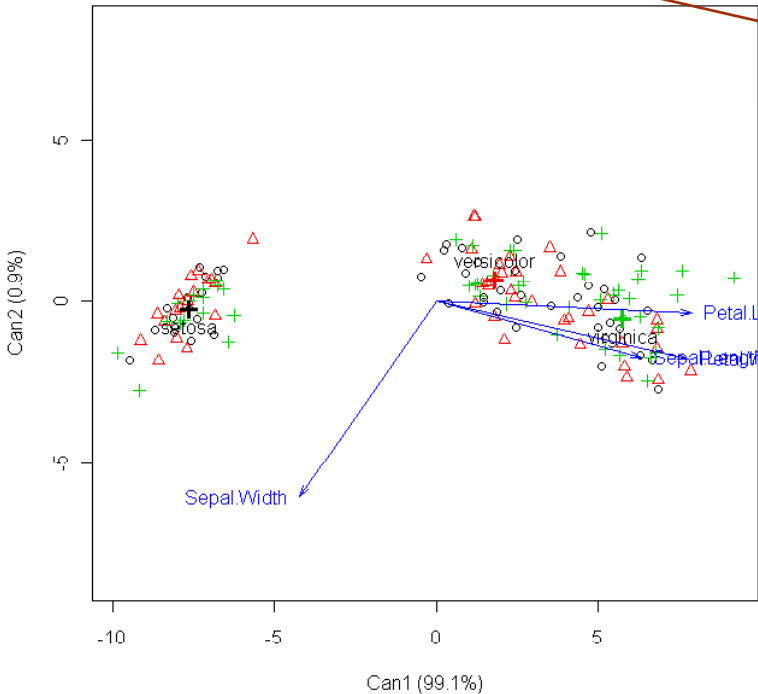
IRIS data 정준상관분석 결과

결과1

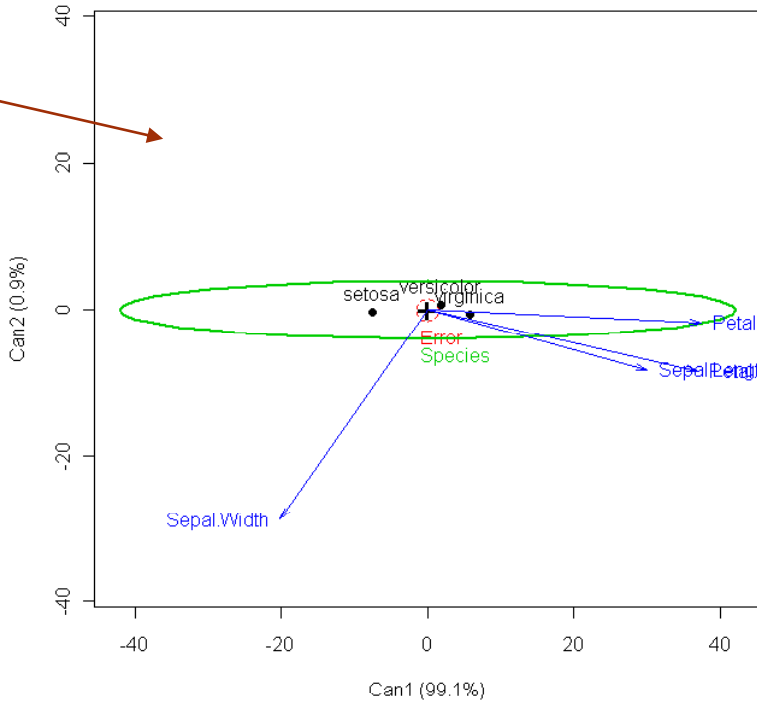
```

> library(candisc)
> iris.mod=lm(cbind(Petal.Length, Sepal.Length,
+ Petal.Width, Sepal.Width) ~ Species, data=iris)
> iris.can=candisc(iris.mod, data=iris)
> plot(iris.can)
Vector scale factor set to 8
> heplot(iris.can)
Vector scale factor set to 38

```



결과2



<http://wolfpack.hnu.ac.kr>



Logistic 판별분석

■ 개념

- 종속변수가 이진형(binary)이거나 순서형(ordinal)인 경우 사용되는 회귀분석

■ 종속변수 Binary: Logit 모형

$$y_i = f(x) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i$$

- 종속변수 측정치 $Y_i=0$ (실패), 1 (성공)
- $p=P(Y_i=1)$
- ODDS ratio(오즈비)= $p/(1-p)$
- Ln(odds)를 종속변수로 사용하여 일반 회귀분석 실시

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

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

■ 회귀계수 B의 의미

- 회귀계수의 부호는 성공확률(p) 증감과 일치
- EXP(회귀계수)는 설명변수가 한 단위 증가할 때 odds ratio에 미치는 영향(multiplication)이 된다.

$$\frac{p_i}{1-p_i} = (e^{\hat{\alpha}})(e^{\hat{\beta}_1})^{x_{1i}} \dots (e^{\hat{\beta}_p})^{x_{pi}}$$

■ 판별분석에 활용 (페이지 229)

- 종속변수 집단에 속할 사후확률이 집단에 속할 확률
- Specification, Sensitivity (event correction)

■ 장점

- 판별변수에 대한 유의성 검정 편리
- 판별에 영향 정도 비교 가능
- 판별변수로 지시변수 사용 가능

■ 종속변수 Ordinal: Logistic 모형

- k는 종속변수 집단 최대 값, 0, 1, 2, ..., k

$$\log it(p_1) = \ln\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \underline{\beta}' \underline{X}$$

$$\log it(p_1 + p_2) = \ln\left(\frac{p_1 + p_2}{1-p_1-p_2}\right) = \beta_0 + \underline{\beta}' \underline{X}$$

...

$$\log it(p_1 + p_2 + \dots + p_k) = \ln\left(\frac{p_1 + p_2 + \dots + p_k}{1-(p_1 + p_2 + \dots + p_k)}\right) = \beta_0 + \underline{\beta}' \underline{X}$$



Logistic 판별분석 예제

http://wolfpack.hnu.ac.kr

TURKEY.txt

```
> # Logistic Regression
> fit=glm(type~uln+tin+car,
+ data=tk,family=binomial())
> summary(fit) # display results
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -53.99740   32.82646  -1.645   0.100 .
uln          0.01642    0.21538   0.076   0.939
tin          0.60772    0.31800   1.911   0.056 .
car         -0.04644    0.03435  -1.352   0.176
```

- 유의하지 않은 ULN 제외

```
> # Logistic Regression
> fit=glm(type~tin+car,
+ data=tk,family=binomial())
> summary(fit) # display results
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -54.95184   30.88457  -1.779   0.07520 .
tin          0.62490    0.23322   2.679   0.00737 **
car         -0.04540    0.03125  -1.453   0.14626
---
AIC: 18.746
```

- 알파벳이 큰 Wild가 event에 할당 $y=P(\text{Wild})$

회귀계수 신뢰구간

```
> confint(fit) # 95% CI for the coefficients
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) -136.1757281 -4.873863675
tin           0.2887465  1.250926848
car          -0.1263470  0.006628248

> exp(confint(fit)) # 95% CI for expo
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) 7.238236e-60 0.007643775
tin          1.334753e+00 3.493579476
car          8.813090e-01 1.006650263
```

종속변수 사후확률

```
> predict(fit, type="response") # pred
      1          2          3
9.100764e-01 9.929691e-01 9.847129e-01
      7          8          9
7.052489e-01 9.998387e-01 9.996001e-01
```

집단 분류 (사후확률 이용)

```
> yhat$class[yhat$fit.fitted.values>0.5]="Wild"
> yhat=data.frame(fit$fitted.values)
> yhat$class[yhat$fit.fitted.values>=0.5]="Wild"
> yhat$class[yhat$fit.fitted.values<0.5]="Domestic"
> yhat
  fit.fitted.values  class
1  9.710665e-01    Wild
2  9.991106e-01    Wild
3  9.996327e-01    Wild
4  9.999999e-01    Wild
```



K-nearest Neighbor DA (페이지 205)

■ 비모수 방법

- 판별변수가 다변량 정규분포를 따르지 않을 경우
 - (1) 분류하려는 개체와 Mahalanobis 거리가 가장 가까운 개체를 구하고 그 개체가 속한 집단으로 분류한다.
 - (2) 만약 거리가 같은 개체가 2개인 경우 동일 집단이면 그 집단에 분류한다.
 - (3) 2개이면서 그 개체의 집단이 동일하지 않으면 그 다음 가까운 개체의 집단을 조사하여 3개의 개체 중 많이 속한 집단으로 분류한다. 여기서 k nearest neighbor 의미는 Mahalanobis 거리가 가장 가까운 개체 k개를 고려하여 그 k개 개체의 군집 중 가장 많은 수를 차지하는 군집에 분류하게 된다. 다음 프로그램 거리가 가장 가까운 3개의 개체들의 집단을 조사하여 가장 많은 집단으로 분류하는 방법이다.

■ 새로운 접근방법

- 판별 변수(측정 변수)가 이산형, 순서형 분류형, Binary인 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. Hartigan이 개발한 CHAID(Chi-square Automatic Interaction Detector)가 있다. 이 방법은 현재 Data Mining 기법으로 가장 많이 이용되고 있다. SPSS에는 ANSWER tree TOOL에 속해 있다.

■ Term project on DA due 2008.12.08 (Mon)

- 다음 방법에 의해 판별분석을 실시하고 오분류 표 작성
 - Fisher 판별분석 (with, without 판별변수 선택)
 - Logistic 판별분석 (유의한 변수) 이진형
 - 정준 판별분석
- 최적 판별분석 방법으로 개체 판별한 결과 해석

