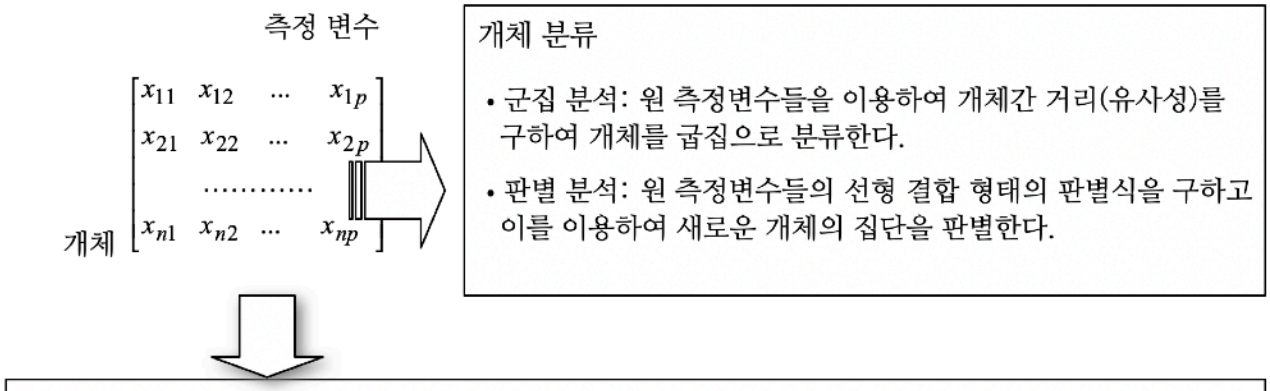
판별분석 개요

1) 다변량 분석 비교

판별 분석(Discriminant Analysis)은 군집 분석(Clustering Analysis)과 함께 개체 유도 기법(individual directed techniques)이다.

변수 유도 기법 (Variable-directed techniques)은 p 개의 원 변수 $(x_1, x_2, ..., x_p)$ 들의 상관 관계(공분산 행렬이나 상관계수 행렬)를 이용하여 주성분 변수(원 변수의 선형결합)를 도출하여 80% 규칙에 의해 p차원을 2~4개의 주성분 차원으로 축약하는 주성분 분석과 원 변수를 요인(factor)이라는 잠재 변인을 이용하여 상호 배타적으로 그룹화하는 방법인 요인분석으로 나눌 수 있다.

판별 분석(Discriminant Analysis)은 군집 분석(Clustering Analysis)과 함께 개체들에 대해 측정된 특성(변수) 값을 이용하여 개체를 판별하는 식을 유도하여 새로운 개체의 집단을 판별하거나 개체의 유사성을 계산하여 유사한 개체끼리 군집화 하는 개체 유도 기법(individual directed techniques)이다.



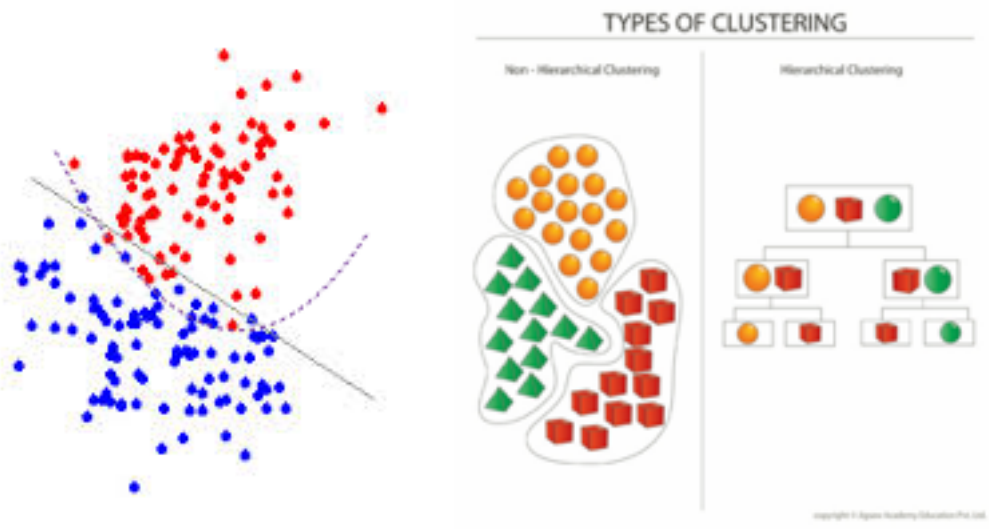
2) 군집분석과 판별분석 차이

졸업생의 취업여부에 영향을 주는 변수로 평점, 비만도(=키/몸무게: 외모척도), 영어 성적, 자격증 개수, 원서 지원 회수, 가족 총 연 소득 (재정 능력), 친구 수(사교력 측정)를 고려하였다. 졸업생 중 6개월 전에 취업한 졸업생 30명, 취업하지 못한 졸업생 20명을 무작위 추출(취업율이 60%) 하여 7개 항목을 조사하였다.

측정된 7개 변수를 이용하여 취업여부(2개의 집단)를 판별하는 판별규칙(판별함수)을 도출하여 새로운 개체(졸업 예정자)에 대하여 졸업 후 6개월 이전에 취업여부를 확률적으로 측정하는 방법을 판별분석이라 한다. 판별분석은 분석 전 개체를 분류하는 집단(그룹) 정보가 있다.

H대학교 비즈니스 통계학과 졸업 예정자를 대상으로 위의 7개의 변수(학생들의 취업 경쟁력)들을 측정 조사하였다고 하자. 취업 경쟁력 관점에서 졸업예정자들의 유사성을 측정하여(일반적으로 유크리디안 거리) 유사성이 높은 개체끼리 상호 배타적으로 분류하는 것을 군집분석이라 한다. 군집분석의 핵심적 단계는 유사성 계산, 유사성에 의한 집단으로 분류해 가는 기준, 군집의 개수 결정, 개체 군집에 대한 적절한 이름 부여이다. 군집은 판별분석의 집단과 유사한 개념이다.

좌측 판별분석은 2개의 집단을 판별하는 직선 판별식, 2차 판별식을 그린 예제이며, 우측 군집분석은 개체를 3개의 군집으로 나눈 것으로 비계층적 방법과 계층적 방법을 적용한 예이다.



(구글 서치 이미지)

판별분석 군집분석

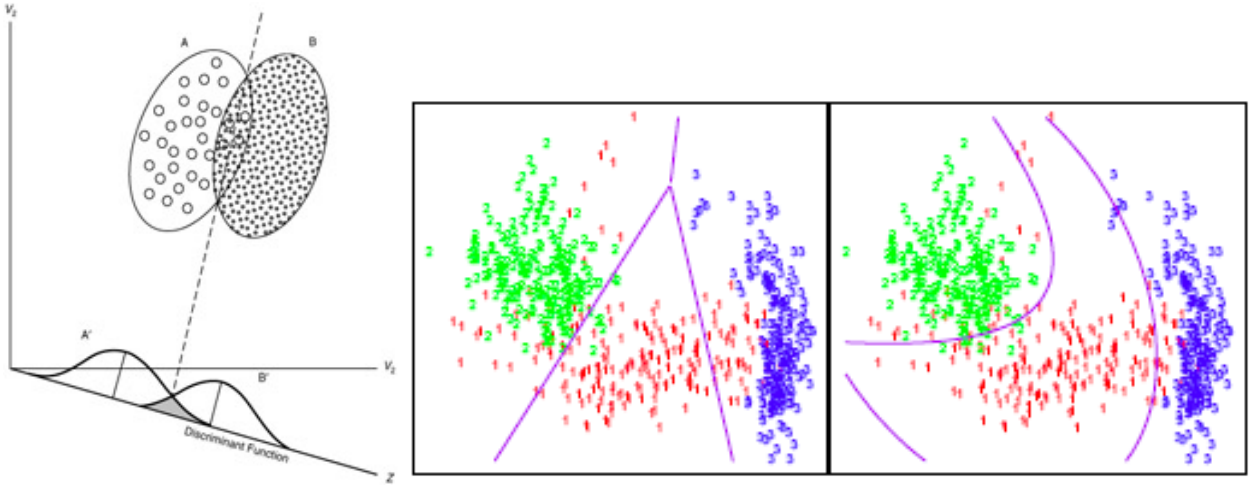
군집분석에서는 개체의 그룹에 대한 정보 없이 (사전에는 ●, △, ☒구분이 없이 동일하나 분석 후 나누어짐) 유사성이 가까운 개체들끼리 계층적으로 묶어 가거나 군집의 개수를 정하여 군집의 중심점을 이용하여 개체를 군집화 하는 방법이다. 판별분석은 자료 수집 시 이미 그룹이 나누어져 있어(●, ●) 이를 가장 잘 판별하는 판별규칙을 도출하여 새로운 개체의 군집을 판별하는 방법이다. 곡선의 판별규칙을 도출하는 것은 쉽지 않다.

3) 판별규칙

판별규칙은 모집단 그룹을 정확하게(오분류 비율 최소화) 하는 '판별' 변수의 함수식으로 판별규칙을 설정한다. 아래 그림은 인터넷 구글링 결과 얻은 그림이다.

[왼쪽] 2개 모집단, 판별변수 (y, z)의 함수식으로 점선 직선=판별규칙을 얻었음. 아래 히스토그램의 빗금친 부분 중 (B 모집단의 판별규칙 왼쪽 꼬리 빗금친 확률)=B 모집단에 속한 개체를 판별규칙에 의해 A집단으로 분류한 오분류 확률, (A 모집단의 판별규칙 오른쪽 꼬리 빗금친 확률)=A 모집단에 속한 개체를 판별규칙에 의해 B집단으로 분류한 오분류 확률

[오른쪽] 왼쪽은 직선 판별규칙이고 오른쪽은 다항식 판별규칙임. 현행 산점도를 보면 다항식 판별규칙이 오분류가 적어 보임. 그러나 실제 이차식까지만 판별규칙이 가능, 실제 판별규칙의 선택은 오분류 결과 활용



4) 오분류 misclassification ratio

마취과 의사가 마취 양을 결정할 때 (나이, 혈압, 맥박, 몸무게)가 영향을 준다고 하자. 이 4개의 변수만으로 마취량을 결정할 때 문제는 없을까? 위의 4개 변수로 마취안전여부를 판단하는 판별규칙을 만들었다고 하자. 이 판별규칙의 정확도? 어떻게 측정할까?

분류 정의

(정분류) = 판별규칙에 의해 (성공 개체를 성공), (실패 개체를 실패)로 분류하는 결과

(오분류) = 판별규칙에 의해 (실패 개체를 성공), (성공 개체를 실패)로 분류하는 결과

판별결과=>	마취안전	마취위험
마취안전	정분류 (a)	(1)오분류 (b)
마취위험	(2)오분류 (c)	정분류 (d)

통계적 가설

- 귀무가설 : state of nature, 환자는 마취에 안전
- 대립가설 : 환자는 마취에 위험
- 1종 오류 : (b), 2종 오류 : (c),
- 검정력 : 1-(c) = 마취 위험 환자를 위험하다고 진단

용어정의

Positive 양성 - 감염, 마취위험, 불량

Negative 음성 - 정상, 마취위험 없음, 정품 = 데이터 총 수 = $(a+b+c+d)$

정분류 비율 = $\frac{(a+d)}{(a+b+c+d)}$, 오분류 비율 = $\frac{(b+c)}{(a+b+c+d)}$

4) 판별규칙 결정 방법

(판별규칙 1) a=45, b=5, c=0, d=50 -> 오분류비율=95%

(판별규칙 2) a=50, b=0, c=5, d=45 -> 오분류비율=95%

어느 규칙을 선택할 것인가? 오분류비율만 보면 동일하다. 그러나 판별규칙2의 비용은 너무 크다. 마취 위험한 환자를 안전하다고 결정하면 수술 중 환자가 위험에 빠질 가능성이 높아 기회비용이 높다. 그래서 판별규칙을 결정할 때는 오분류비율도 고려하지만 오분류 내용에 의한 비용문제도 고려해서 결정한다.

5) 오분류 추정방법

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. 그러나 대부분의 소프트웨어는 이를 사용한다. 빅데이터에서는 train(학습), test(감증) 데이터로 나누어 판별규칙 검증한다.

테스트 데이터 이용

데이터를 양분하여 한 개체 그룹으로부터 판별식을 유도하고, 이 판별식을 사용하여 다른 그룹의 개체를 분류하여 오분류 비율을 추정한다. 표본 자료의 1/2만 사용하여 판별식을 구하므로 모집단 분류에 적합한

판별식을 얻을 가능성이 낮고 데이터를 많이 수집해야 한다는 단점으로 인하여 이 방법 역시 사용 빈도가 낮았지만 빅데이터에서는 가능하다. 가장 많이 사용.

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.

6) Positive, Negative 개념

환자의 유병여부를 진단하는 진단키트(거짓말 탐지기)의 테스트 결과

양성 positive ={병이 있는 사람에게 병이 있다고 판단} ={거짓말하는 사람을 거짓말 하고 있다고 진단}

음성 negative ={건강한 사람을 건강하다고 판단} ={진실한 사람을 진실하다고 진단}

진단결과=>	양성 <=> H0 기각	음성 <=> H0 채택
H1: 가설 true 양성(유병, 거짓말)	참양성 true positive TP 검정력 $1-\beta$	거짓음성 false negative FN 2종 오류 β
H0: 귀무가설 true 음성(건강, 진실)	거짓양성 false positive FP (유의수준) α	참음성 true negative TN

민감도 sensitivity, true positive

$$Sensitivity = \frac{TP}{TP+FN}$$: 유병(양성)인 사람을 유병이라고 정확한 진단 확률

특이도 specificity, true negative

$$Specificity = \frac{TN}{FP+TN}$$: 건강(음성)한 사람을 건강하다고 정확한 진단 확률

양성예측도 positive predictive value

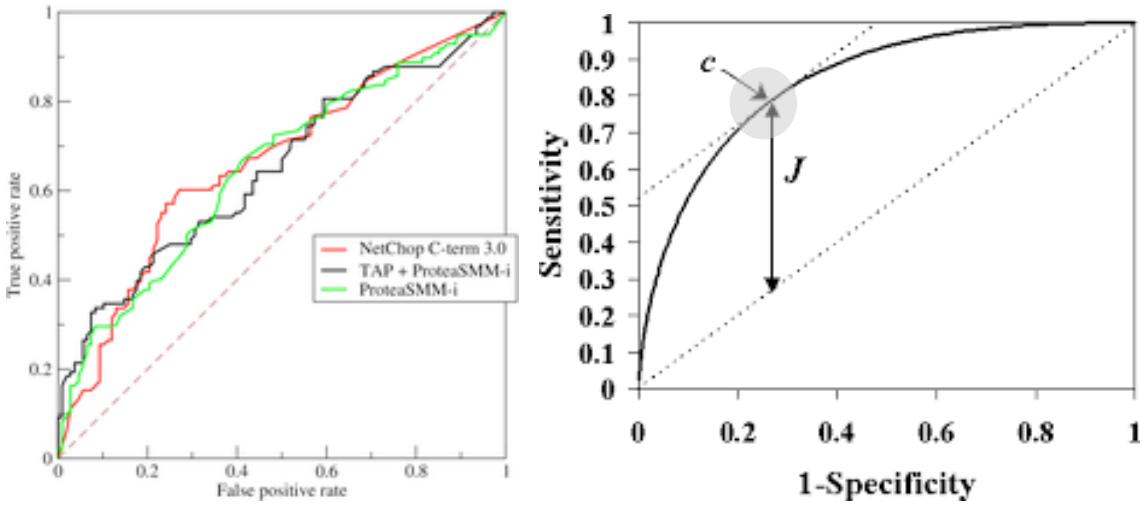
$$PPV = \frac{TP}{TP+FP}$$: 유병(양성)이라 진단한 사람 중 유병인 사람의 확률

음성예측도 negative predictive value

$$NPV = \frac{TN}{TN + FN}$$: 건강(음성)이라 진단한 사람 중 건강한 사람의 확률

민감도, 특이도는 실험,대조군 비교연구에서 예측도는 유병률을 알 수 있는 횡단 연구에 활용된다.

Reciever Operating Curve 분석| (위키피디아)



- x-축을 (1-특이도), y축을 민감도로 하여 그린 이차원 그래프를 활용하여 (양성, 음성)을 판단하는 기준값을 결정하는 기법으로 의학적 분야에 사용
- 진단방법의 기준값 cur off value 을 결정
- 원점은 특이도가 1이므로 음성을 음성을 완벽하게 진단하는 기준값으로 양성 확률 기준값이 1에 해당
- (1, 0)은 민감도 1, 특이도 0이고 (1,1)은 민감도1, 특이도 1이므로 완벽한 진단임
- 곡선 아래 면적을 AUC(area under the curve)라 하고 이 값이 가장 큰 방법이 가장 적절한 진단 방법임
- 45도 대각선은 "양성" 진단을 무작위로 진단하였다는 것을 의미함, 그러므로 진단방법이 무작위 이상이어야 적절하므로 AUC 면적은 1/2보다 커애 함
- 진단방법 결정 : AUC 값이 가장 큰 진단 방법을 결정하면 된다. 위의 그림에서는 붉은 색 곡선을 그린 진단방법이 적절함
- 기준값 결정방법 : Youden J 통계량이 가장 큰 곳의 값을 기준 값으로 함 - 다음 그림의 원의 c값을 만들어내 기준값을 선택하면 된다.

예제 데이터

1) LPGA 2008 (모집단 2개)

LPGA2008.csv : 미국 LPGA 골프선수들의 능력 측정 변수 6개 (평균비거리(+), 페어웨이 안착율(+), 그린 적중률(+), 평균버팅 수(-), 샌드회수(-), 샌드세이브(+)), 상금, 출전 라운드 수를 조사한 데이터이다. (측정변수 옆 +의 의미는 변수 값이 높을수록 선수능력이 높다는 의미, -는 반대 개념)

상금 순위 40위 선수를 '상위', 40위 미만인 선수를 '하위' 2개 그룹으로 나누고 '상금순위그룹' 변수를 만들고 상금 순위를 판별하는 규칙을 선수 능력변수로 만들고 규칙의 정확성을 파악하자.

```
lpga<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/
lpga2008.csv', fileEncoding = "utf-8")
names(lpga)
```

```
> names(lpga)
[1] "골퍼"              "평균_비거리"      "페어웨이_안착율" "그린_적중률"      "평균_퍼팅수"      "샌드_회수"
[7] "샌드_세이브"      "상금"             "참가_라운드수"
```

상금 크기로 순위를 만들고 (rank(변수) 함수는 변수 크기 순으로 개체를 순위를 부여함. -변수는 역순위 부여) 순위가 40위 이상인 선수는 '상', 그렇지 않은 선수는 '하'

```
lpga$상금그룹[rank(-lpga$상금)<=40]<-'상'
lpga$상금그룹[rank(-lpga$상금)>40]<-'하'
head(lpga$상금그룹); ; head(lpga$상금그룹)
```

```
> head(lpga$상금); head(lpga$상금그룹)
[1]  6063 19343  1873  1212  2555  2282
[1] "하" "상" "하" "하" "하" "하"
```

2) IRIS 데이터 (모집단 그룹 3개)

다변량 데이터셋, 영국 통계학자, 생물학자 Ronald Fisher(1936)

n=150 samples 분꽃 3종 (Iris setosa, Iris virginica and Iris versicolor)

4개 척도 변수 Sepal (꽃받침) 넓이, 길이 / Petal 꽃잎 넓이, 길이를 이용하여 분꽃 종을 판별하는 규칙을 만들고

모집단 집단이 2개, 판별분석

1) 정의

판별 측정변수 벡터 : $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \sim (\underline{\mu}, \Sigma)$

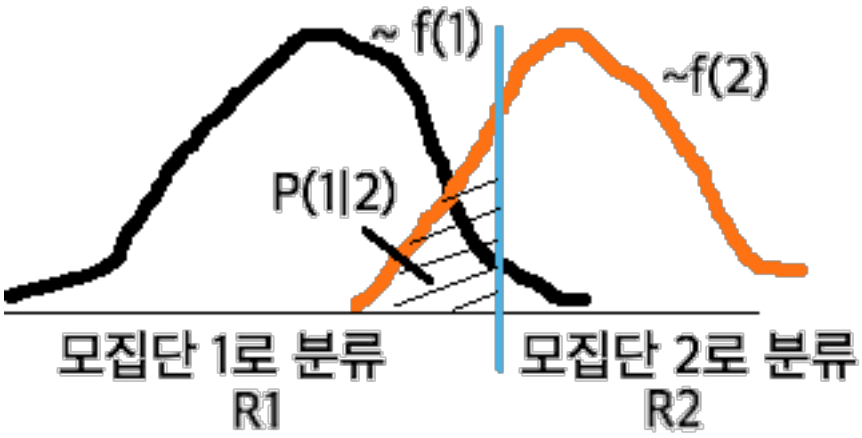
모집단 1 : $\pi_1 \sim f_1(\underline{x}) \sim MN(\underline{\mu}_1, \Sigma_1)$, 모집단 2 : $\pi_2 \sim f_2(\underline{x}) \sim MN(\underline{\mu}_2, \Sigma_2)$

각 모집단으로부터 (n_1, n_2) 크기의 표본 개체를 추출하여 판별 측정변수의 관측값을 조사했다고 하자.

각 표본의 평균벡터를 $(\underline{x}_1, \underline{x}_2)$, 공분산행렬을 (S_1, S_2) 이라 하자.

기호 notation

- $P(\pi_i)$ = 모집단 i의 사전확률 ($\pi_1 + \pi_2 = 1$)
- $P(X \in R_i \mid \pi_j)$ = 모집단 j에 속한 개체가 판별식에 의해 모집단 i로 판별할 확률 , $i = j$이면 정분류, $i \neq j$이면 오분류
- 비용 $c(j \mid i)$ = 모집단 i에 속한 개체를 모집단 j로 판별하여 발생하는 비용. $i = j$ 이면 비용 0임
- $P(\pi_1)P(X \in R_1 \mid \pi_1)$ = 모집단 1에서 추출된 개체를 모집단 1로 정분류
- $P(\pi_2)P(X \in R_2 \mid \pi_2)$ = 모집단 2에서 추출된 개체를 모집단 2로 정분류
- $P(\pi_1)P(X \in R_2 \mid \pi_1)$ = 모집단 1에서 추출된 개체를 모집단 2로 오분류
- $P(\pi_2)P(X \in R_1 \mid \pi_2)$ = 모집단 2에서 추출된 개체를 모집단 1로 오분류



2) 판별규칙

(1) 오분류 기대비용 ECM 최소화

오분류 가대비용 = (오분류확률*비용함수)

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

=> (결과) $R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq (\frac{p_2}{p_1})(\frac{c(1|2)}{c(2|1)})$

$$R_2 : \frac{f_2(\underline{x})}{f_1(\underline{x})} \geq (\frac{p_1}{p_2})(\frac{c(2|1)}{c(1|2)})$$

(만약 비용이 동일하다면 $R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq (\frac{p_2}{p_1})$) (모집단 비율도 동일하다면 $R_1 : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq 1$

(2) 우도함수 likelihood ratio (다변량 정규분포)

모집단이 다변량정규분포를 따른다면, 새로운 개체 $\underline{x}_0$에 대하여

$L(\underline{x}_0; \underline{x}_1, \hat{\Sigma}_1) \geq L(\underline{x}_0; \underline{x}_2, \hat{\Sigma}_2)$ 이 성립하면 모집단 1로 그렇지 않으면 모집단 2로 판별한다.

(3) Fisher 선형 linear 판별

다변량 정규분포의 두 모집단이 동일한 공분산(Σ)을 갖는다면 다음과 같이 통합분산에 의해 추정된다.

통합 공분산 : $\hat{\Sigma} = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}$

그리고 우도함수식은 다음의 판별규칙으로 간편화 된다.

$$(\underline{x}_1 - \underline{x}_2)'\hat{\Sigma}^{-1}\underline{x}_0 - \frac{1}{2}(\underline{x}_1 - \underline{x}_2)'\hat{\Sigma}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \geq c \quad , \quad c = \ln[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}]$$

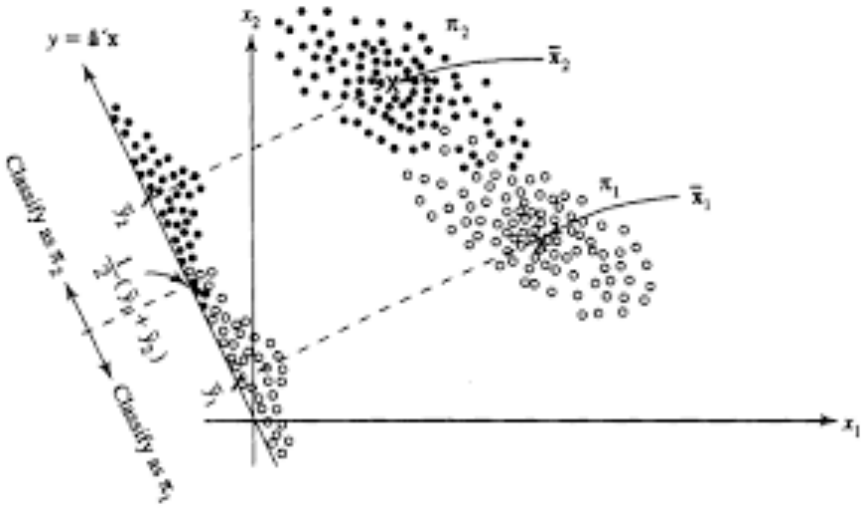
만약 비용이 동일하고 모집단 사전확률이 동일하면 $c = 0$

새로운 개체 $\underline{x}_0$가 위의 식을 만족하면 모집단 1, 그렇지 않으면 모집단 2로 판별한다.

(4) 사후 posterior 확률 이용

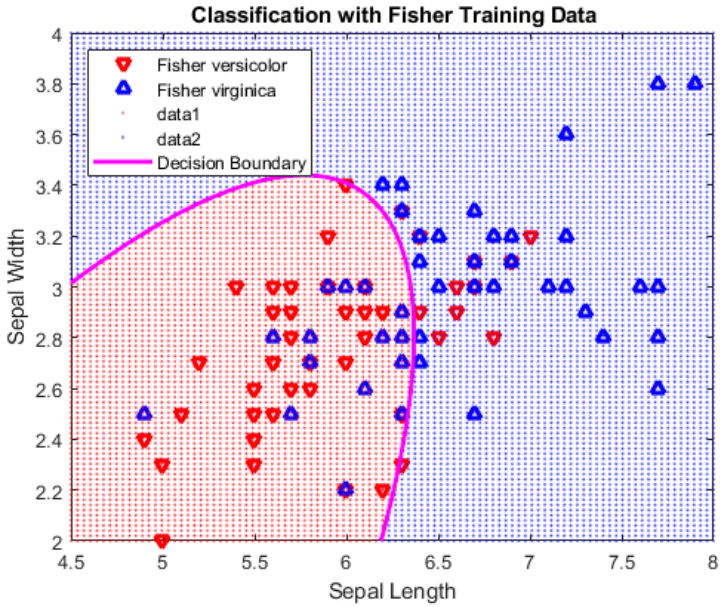
관측치 $x_0 = (x_{10}, x_{20}, \ldots, x_{p0})'$(판별변수 p개인 경우) 모집단($\pi_k$, $k = 1,2$)에 속할 사후 확률 계산 : $\delta_k(x_0) = P(x_0 \hat{\subset} \pi_k) = \underline{x}\hat{\Sigma}^{-1} - \frac{1}{2}\hat{\mu}_k\hat{\Sigma}^{-1} - \hat{\mu}_k + ln(\hat{\pi}_k)$

$P(\pi_1 \supseteq \underline{x}_0 \,|\, data) \geq P(\pi_2 \supseteq \underline{x}_0 \,|\, data)$ 이면 새로운 개체 $\underline{x}_0$는 모집단 1로 판별한다.



(5) 두 모집단의 공분산이 동일하지 않은 경우 Fisher는 Quadratic 판별규칙 제안

사후확률 계산식 $$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$



3) 실증분석 (모집단 2개 LPGA 상금그룹 '상' '하' 선수 판별규칙)

(1) 산점도 행렬 (모집단 그룹별)

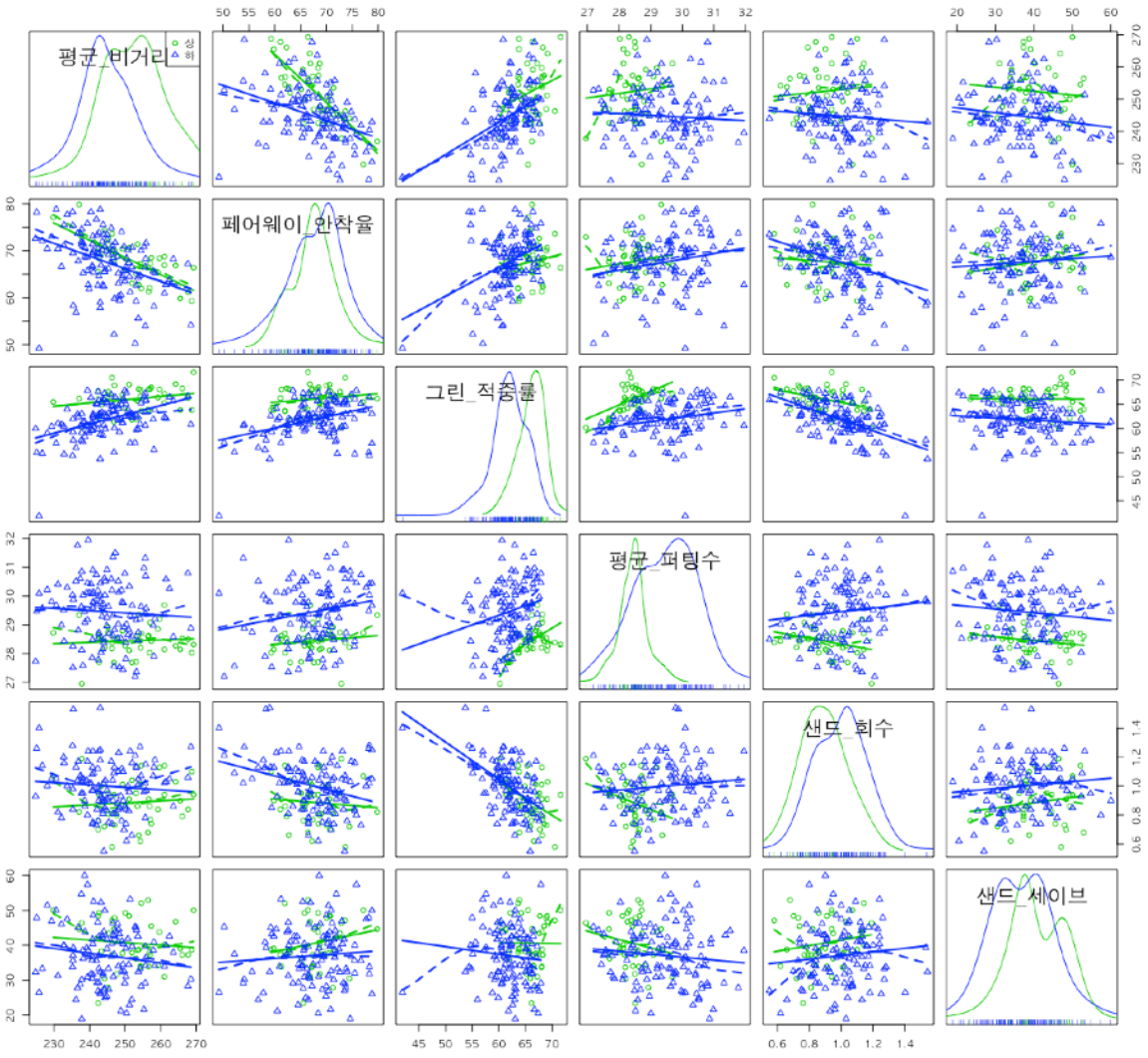
모집단 그룹별 산점도, 히스토그램을 활용하여 모집단 그룹 판별에 영향을 미치는 변수를 시각적으로 판단 가능하다.

```
library(car)
scatterplotMatrix(~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수
+샌드_세이브|상금그룹,data=lpga,col=c(3,4))
```

초록색 (상위그룹) 판별 가능한 변수 (시각적 판단) - 초록색(상 집단), 파랑색(하집단)

히스트로그램 : 오른쪽에 있는 분포 집단이 해당변수 큰 집단, 산점도는 상하, 좌우 구분

평균비거리 (+), 그린적중률(+), 평균퍼팅수(-)



(2) 모집단 등분산 가정 검정

Fisher 판별규칙 중 선형(등분간), 이차식(이분산) 방법 적용 판단

- 귀무가설 : 모집단 그룹간 공분산은 동일하다.
- 대립가설 : 모집단 그룹간 공분산은 상이하다.
- 검정통계량 : Box's (1949) M-test for homogeneity of covariance matrices obtained from multivariate normal data according to one or more classification factors.

```
install.packages('heplots')
library(heplots)
boxM(lpga[,2:7],lpga$상금그룹)
```

유의확률이 0.001보다 작으므로 '등분산' 귀무가설이 기각되어 Fisher 이차 판별 규칙 이용하는 것이 적절하다.

```
> boxM(lpga[,2:7],lpga$상금그룹)

        Box's M-test for Homogeneity of Covariance Matrices

data:  lpga[, 2:7]
Chi-Sq (approx.) = 59.804, df = 21, p-value = 1.367e-05
```

(****) Fisher 선형 판별 (등분산 가정 만족) 무너졌으므로(2) 결과) '이차판별 적정' (예제활용)

```
library(MASS)
lpga.lda<-lda(상금그룹~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,data=lpga) #판별규칙
lpga.lda #판별규칙 결과 출력
lpga.lda.p<-predict(lpga.lda) #사후확률(posterior), 판별값(x)
head(lpga.lda.p$posterior)
```

사전 확률 : 디폴트는 표본 집단그룹 비율대로 모집단 비율을 가정하게 된다. (디폴트)

```
Prior probabilities of groups:
       상        하
0.2547771 0.7452229
```

```
> dim(lpga)
[1] 157  10
```

(표본 157선수이므로 40/157=0.2548 '상' 모집단 비율로 가정됨)

모집단 그룹의 비율을 사전에 안다면 이를 판별규칙에 활용

만약 모집단 그룹 비율에 대한 사전 정보가 있다면 이를 반영하여 판별규칙을 결정하여야 판별 정확성이 높아진다. 사전확률이 정의하면 판별 규칙도 달라진다. 다음은 2008년 활동 선수가 200명이라면 상위 40위 비율은 0.2이므로 모집단 비율 사전 정보를 설정하여 판별규칙 도출 시 이를 고려하게 하는 것이 적절함

```
lpga.lda<-lda(상금그룹~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,data=lpga, prior=c(0.2,0.8))
```

```
Prior probabilities of groups:
 상  하
0.2 0.8
```

(3) 판별분석 : 등분산 가정이 무너져 이차판별, 그러나 여기서는 선형판별 방법 적용 예제임

```
library(MASS)
lpga.lda<-lda(상금그룹~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,data=lpga)
lpga.lda #판별규칙 결과 출력
```

prior=c(비율1, 비율2) 옵션을 사용하지 않으면 판별규칙의 모집단 그룹 비율은 사용된 표본데이터 비율을 사용 **디폴트 옵션** 사용 (표본이 그룹을 층화변수로 하여 확률비례할당 추출을 가정하거나 모집단 사전 정보가 없는 경우)

```
Prior probabilities of groups:
       상        하
0.2547771 0.7452229
```

판별규칙의 선형계수

LD1=0.0028*(평균_비거리) + 0.0299*(페어웨이_안착율) -0.279*(그린_적중률) + 0.813*(평균_퍼팅수) - 0.857*(샌드_회수) -0.0266*(샌드_세이브) => 각 관측치의 판별값이 계산된다.

```
Coefficients of linear discriminants:
                          LD1
평균_비거리         0.00280365
페어웨이_안착율     0.02992010
그린_적중률        -0.27874740
평균_퍼팅수         0.81317986
샌드_회수          -0.85687947
샌드_세이브        -0.02655564
```

(4) 사후확률 및 판별값

```
lpga.lda.p<-predict(lpga.lda) #사후확률(posterior), 판별값(x)
head(lpga.lda.p$posterior)
lpga.lda.p$class
head(lpga.lda.p$x)
```

도출된 판별규칙에 의해 1번 선수의 '상' 집단에 속할 확률은 0.51, '하' 집단에 속할 확률은 0.49

만약 사후확률 0.5 기준으로 표본 개체(선수)를 분류한다면 1번 선수(안시현)는 상으로 분류 = 오분류, 2번 핼렌은 상으로 정분류, 3번 Dina, 4번 Beth 모두 '하'로 정분류된다.

```
> head(lpga.lda.p$posterior)
          상         하
1 0.50899871 0.4910013
2 0.59300108 0.4069989
3 0.04121127 0.9587887
4 0.03470754 0.9652925
5 0.22497126 0.7750287
6 0.05776803 0.9422320
```

```
> head(lpga[,c(1,8,10)])
                골퍼  상금 상금그룹
1     Ahn, Shi Hyun  6063       하
2 Alfredsson, Helen 19343       상
3 Ammaccapane, Dina  1873       하
4       Bader, Beth  1212       하
```

디폴트 0.5기준으로 추정 사후확률이 큰 모집단 그룹으로 분류하여 $class에 저장되어 있음

```
> lpga.lda.p$class
 [1] 상 상 하 하 하 하 하 하 상 하 하 하 하 하 하 하
[34] 하 하 하 하 하 하 하 하 하 하 하 하 하 하 상 하
[67] 하 하 하 하 하 하 상 하 상 상 상 상 하 하 하 하
```

판별 선형계수에 의해 계산된 관측치의 판별값이 출력된다. 0을 중심으로 멀어질수록 각 집단에 속할 확률이 높아진다. 1번 선수보다 2번 선수의 '상'집단 속할 확률이 높으므로 음의 값이 크다. 3번과 4번 선수의 경우 4번이 '하' 집단에 속할 확률이 크므로 4번 판별값이 3번보다 더 크다.

```
> head(lpga.lda.p$x)
         LD1
1 -1.0456566
2 -1.2172575
3  0.5589827
4  0.6489783
5 -0.4039304
6  0.3799427
```

분류표 산출

```
lpga.lda.ct<-table(lpga$상금그룹,lpga.lda.p$class)
prop.table(lpga.lda.ct,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.lda.ct))) #정분류
```

모집단 '상' 그룹 40명 선수 중 피셔판별규칙에 의해 '상'으로 정분류(31명) 비율은 77.5%, '하'로(9명) 오분류 비율은 22.5%임

모집단 '하' 그룹 117명 선수 중 피셔판별규칙에 의해 '하'으로 정분류(112명) 비율은 95.7%, '하'로(5명) 오분류 비율은 4.2%임

피셔판별규칙은 상위 그룹 선수의 판별력은 77.5%, 하위 그룹 선수의 판별력은 95.7%

```
> lpga.lda.ct

    상  하
  상 31   9
  하  5 112
```

```
> prop.table(lpga.lda.ct,1) #행 퍼센트,분류표

          상         하
  상 0.77500000 0.22500000
  하 0.04273504 0.95726496
```

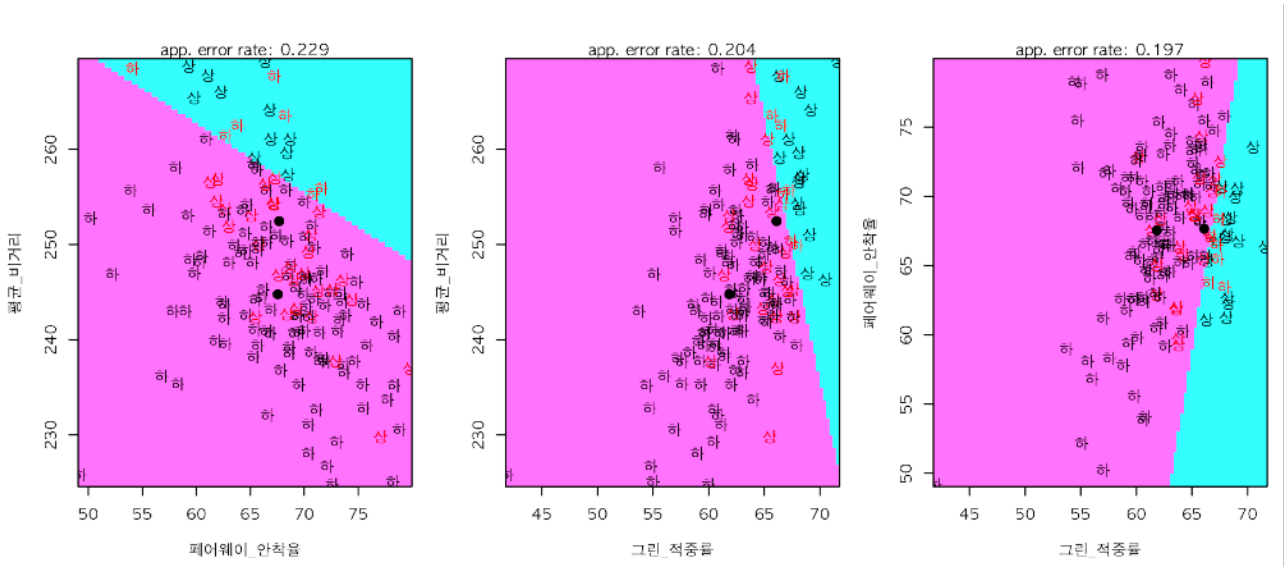
전체 선수 정분류 비율 = (31+112)/157 (91.1%)

```
> sum(diag(prop.table(lpga.lda.ct))) #정분류
[1] 0.910828
```

(5) 시각화 (판별변수 쌍체 산점도, 판별규칙, 오분류 개체 표현)

```
library(klaR)
partimat(as.factor(상금그룹)~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수
+샌드_회수+샌드_세이브,data=lpga,method='lda')
```

모든 판별변수(평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브) 쌍으로 산점도에 원 집단을 개체 레이블로 표시하고 (선형) 판별규칙을 나타낸다. 그리고 판별규칙에 의해 오분류 개체(선수)는 붉은 색으로 레이블 한다. 그리고 오분류 비율을 산점도 제목에 출력한다. 다음 산점도는 총 15개($_6C_2$) 산점도 중 일부이다. 판별변수 (비거리, 페어웨이) 2개만으로 도출된 판별규칙은 색을 구별하는 직선이고 이 판별규칙에 의한 오분류 비율은 22.9%이다. (비거리, 그린적중률) 판별규칙은 오분류 비율은 20.4%, (페어웨이, 그린)의 오분류는 19.7% -> 2개 변수만으로 판별한다면 (페어웨이, 그린)이 가장 좋다.



오분류 개체 판별

```
lpga.lda.result<-cbind(lpga,lpga.lda.p$class)
lpga.lda.result$분류결과[lpga.lda.result[,10]==lpga.lda.result[,11] &
lpga.lda.result[,10]=='상']<-'OO'
lpga.lda.result$분류결과[lpga.lda.result[,10]==lpga.lda.result[,11] &
lpga.lda.result[,10]=='하']<-'XX'
lpga.lda.result$분류결과[lpga.lda.result[,10]!=lpga.lda.result[,11] &
lpga.lda.result[,10]=='상']<-'OX'
lpga.lda.result$분류결과[lpga.lda.result[,10]!=lpga.lda.result[,11] &
lpga.lda.result[,10]=='하']<-'XO'
head(lpga.lda.result)
```

```
> head(lpga.lda.result)
               골퍼 평균_비거리 페어웨이_안착율 그린_적중률 평균_퍼팅수 샌
1     Ahn, Shi Hyun        249.4             64.6         61.2
2 Alfredsson, Helen        253.8             62.7         68.2
3 Ammaccapane, Dina        246.3             70.2         64.6
4       Bader, Beth        249.1             64.1         61.2
5       Bae, Kyeong        244.0             62.4         60.7
6     Baena, Marisa        254.2             64.7         60.9
  상금그룹 lpga.lda.p$class 분류결과
1       하                상       XO
2       상                상       OO
3       하                하       XX
```

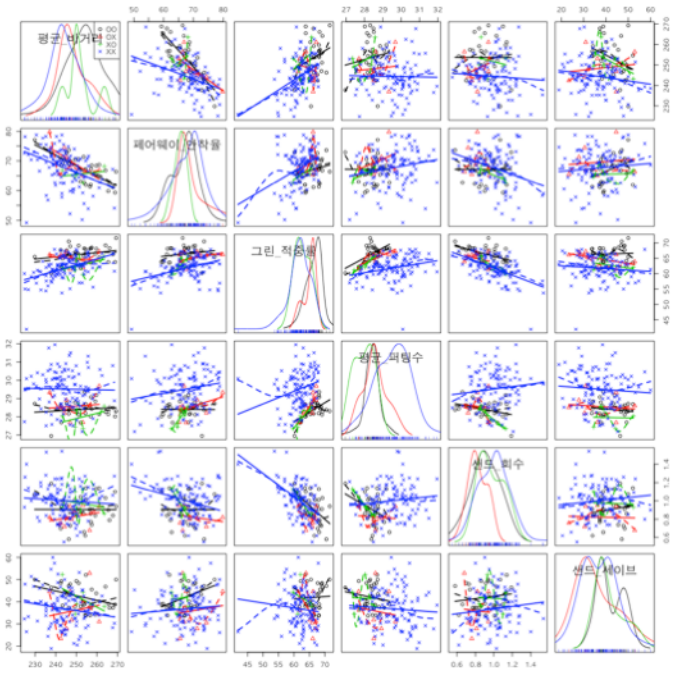
새로 만들어진 '분류결과' 변수는 4개 범주 그룹 변수임

```
> table(lpga.lda.result[,12])

 OO  OX  XO  XX
 31   9   5 112
```

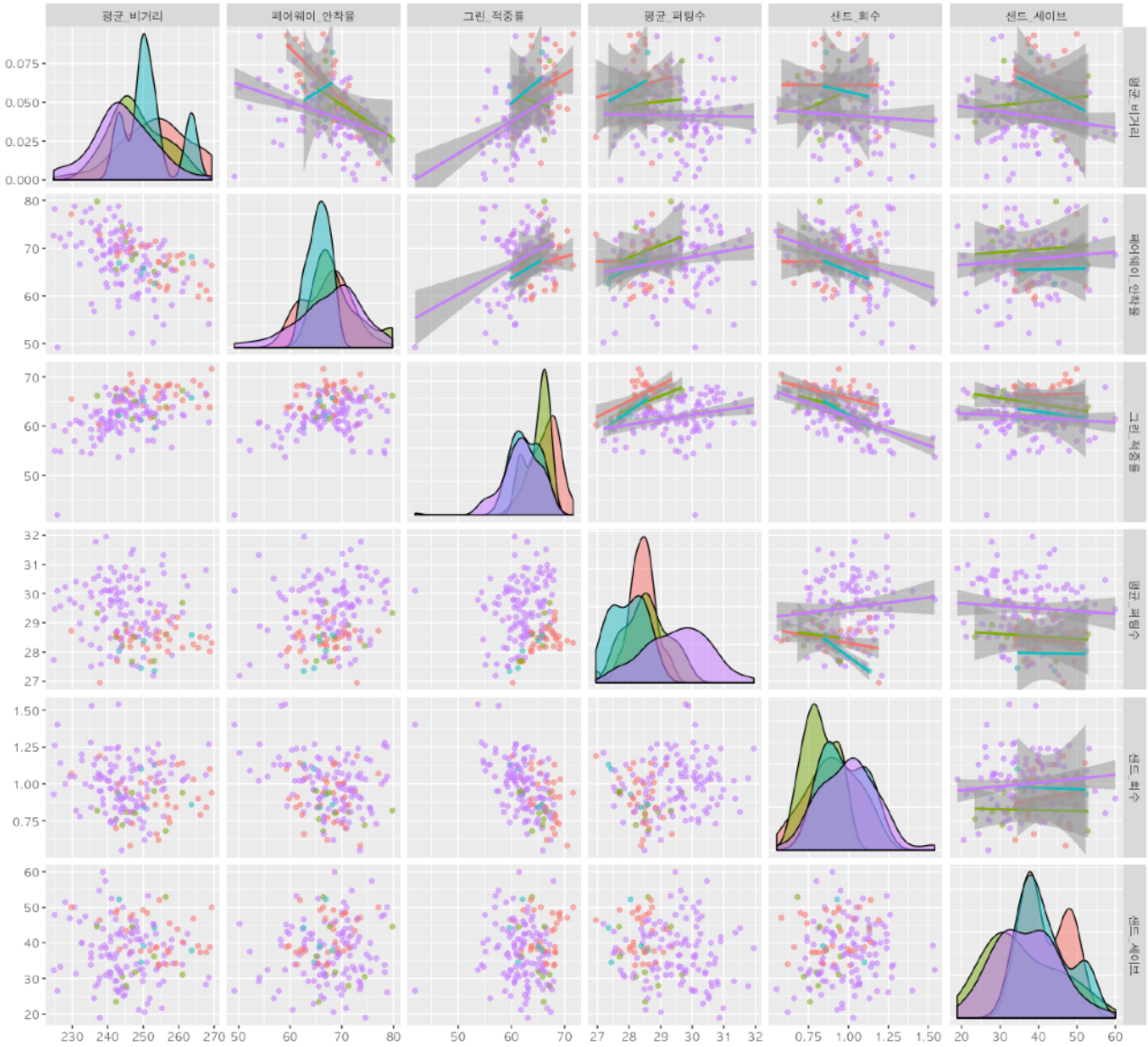
- OO=개체 원집단 '상' - 분류결과 '상', XX=개체 원집단 '하' - 분류결과 '하' - 정분류
- OX=개체 원집단 '상' - 분류결과 '하', OX=개체 원집단 '하' - 분류결과 '상' - 오분류

판별결과 그룹별 산점도

scatterplotMatrix(~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브|분류결과,data=lpga.lda.result,col=c(1:4))



```
library(GGally)
theme_update(text=element_text(family="AppleGothic"))
ggpairs(lpga.lda.result[,2:7],aes(colour=lpga.lda.result$분류결
과,alpha=0.4),upper=list(continuous='smooth'))
```



(6) 판별결과 그룹별 판별변수 평균, 표준편차

```
library(doBy)
summaryBy(평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브~분류결과, data=lpga.lda.result,FUN=c(mean,sd),na.rm=TRUE)
```

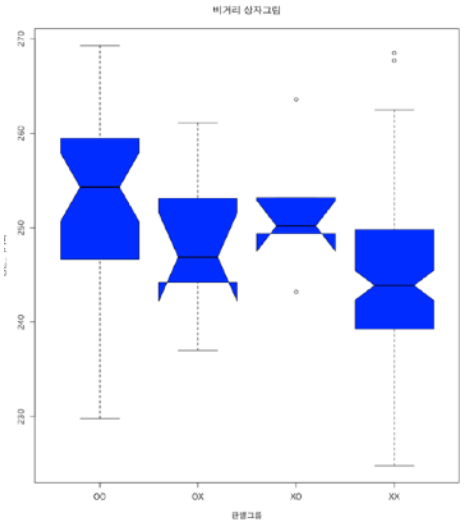
비거리 : 비거리가 큰 선수는 상금 상위 랭킹되므로 평균 253.6 선수는 정분류되지만, 비거리가 다소 낮은 선수는(평균 248.2)는 '하'위 선수로 오분류 되었다. 비거리가 낮은 선수는 하위랭킹 선수이므로 실제 244.4 선수는 '하'위 선수로 정분류 되나, 다소 높은 251.9 부근 선수들의 판별규칙에 의해 상위 선수로 오분류 되었음

	분류결과	평균_비거리.mean	페어웨이_안착율.mean	그린_적중률.mean
1	OO	253.6645		67.12903
2	OX	248.2222		69.52222
3	XO	251.9200		65.60000
4	XX	244.4321		67.63393

	평균_퍼팅수.mean	샌드_회수.mean	샌드_세이브.mean	평균_비거리.s
1	28.40226	0.9054839		41.96129
2	28.57000	0.8244444		35.87778
3	27.95800	0.9720000		41.62000
4	29.51920	1.0008036		36.84732

페어웨이_안착율.sd 그린_적중률.sd 평균_퍼팅수.sd 샌드_회수.sd

```
boxplot(평균_비거리~분류결과,data=lpga.lda.result,notch=TRUE,col='blue',
        main="비거리 상자그림", xlab="판별그룹")
```



(7) 주성분 분석 활용

```
lpga.pca=prcomp(lpga.lda.result[,2:7],scale.=T) #상관계수행렬 이용
summary(lpga.pca) #변동(누적)설명기여율 출력
lpga.pca$rotation #부하 출력
lpga.lda.pca<-cbind(lpga.lda.result,lpga.pca$x)
```

제1 주성분 : 파온 능력(-), 제2주성분 : 장타능력

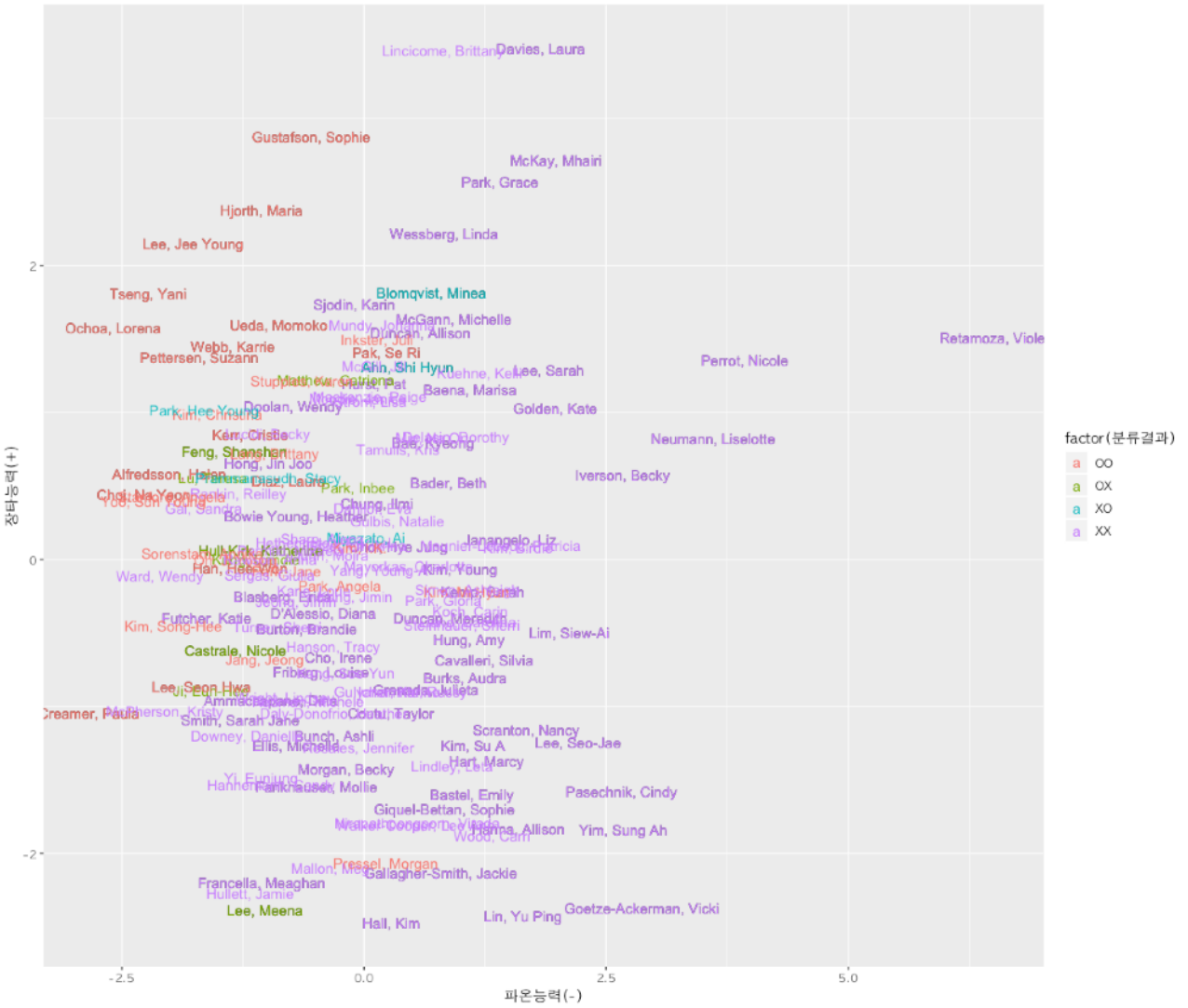
상세 분석 활용 (http://203.247.53.31/Stat_Notes/adv_stat/MDA/MDA_PCA.pdf) 20페이지

```
> summary(lpga.pca) #변동(누적)설명기여율 출력
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6
Standard deviation     1.370 1.239 1.078 0.914 0.6593 0.3921
Proportion of Variance 0.313 0.256 0.194 0.139 0.0724 0.0256
Cumulative Proportion  0.313 0.569 0.763 0.902 0.9744 1.0000
> lpga.pca$rotation #부하 출력
                      PC1       PC2       PC3        PC4      PC5       PC6
평균_비거리      -0.39976  0.57931 -0.15475  0.321000  0.2122 -0.57670
페어웨이_안착율 -0.19523 -0.70654  0.12124 -0.009798  0.5203 -0.42097
그린_적중률      -0.66579 -0.07057 -0.10126  0.277574  0.2082  0.64892
평균_퍼팅수       0.14478 -0.35862 -0.59873  0.582617 -0.3738 -0.11317
샌드_회수         0.58124  0.16983 -0.03577  0.387920  0.6534  0.23363
샌드_세이브       0.00242 -0.05264  0.76899  0.574362 -0.2725 -0.04148
```

```
library(ggplot2)
ggplot(lpga.lda.pca,aes(PC1,PC2,label=골퍼)) +geom_text(check_overlap =
TRUE)+geom_text(aes(colour = factor(분류결과))) +xlab('파온능력(-)')+ylab('장타능
력(+)'); subset(lpga.lda.pca,분류결과=='OX')[,c(1,12)]
```

```
> subset(lpga.lda.pca,분류결과=='OX')[,c(1,12)]
                  골퍼 분류결과
14      Castrale, Nicole       OX
35        Feng, Shanshan       OX
59  Hull-Kirk, Katherine       OX
68           Ji, Eun-Hee       OX
83          Kung, Candie       OX
86            Lee, Meena       OX
95            Lu, Teresa       OX
99     Matthew, Catriona       OX
120          Park, Inbee       OX
```

상위랭크 선수 -> 하위집단 오분류



(8) 새로운 개체 선수 판별규칙에 의한 사후확률 추정

```
predict(lpga.lda,newdata=data.frame(평균_비거리=260,페어웨이_안착율=70,그린_적중률=65,평균_퍼팅수=28,샌드_회수=1.5,샌드_세이브=40))$posterior
```

```
> predict(lpga.lda,newdata=da
     상     하
1 0.861 0.139
```

비거리 260, 페어웨이 70, …, 샌드세이브=40 인 New 선수는 상위 랭크 가능성은 86.1%, 하위랭크 가능성은 13.9%로 판별된다.

실증분석2 Fisher 이차 판별 (등분산 가정 파괴 시) 실제 적절한 방법

```
lpga<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/
lpga2008.csv', fileEncoding = "utf-8")
lpga$상금그룹[rank(-lpga$상금)<=40]<-'상'
lpga$상금그룹[rank(-lpga$상금)>40]<-'하'

library(MASS)
lpga.qda<-qda(상금그룹~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌
드_회수+샌드_세이브,data=lpga)

lpga.qda.p<-predict(lpga.qda) #사후확률(posterior), 판별값(x)
```

이차 판별식이므로 판별규칙의 선형계수 없음 `Coefficients of linear discriminants:`

분류표 산출

```
lpga.qda.ct<-table(lpga$상금그룹,lpga.qda.p$class)
lpga.qda.ct
prop.table(lpga.qda.ct,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.qda.ct))) #정분류
```

모집단 '상' 그룹 40명 선수 중 피셔판별규칙에 의해 '상'으로 정분류(36명) 비율은 90%(선형판별=77.5%), '하'로(4명) 오분류 비율은 10%(선형판별=22.5%)임

모집단 '하' 그룹 117명 선수 중 피셔판별규칙에 의해 '하'으로 정분류(110명) 비율은 94%(선형판별=95.7%, '하'로(7명) 오분류 비율은 6%(선형판별=4.2%)임

이차 판별규칙이 선형판별에 비해 상위랭크 선수 정분류 비율은 12.5% 높고, 하위랜크 선수 정분류 비율은 1.7% 낮다.

```
> lpga.qda.ct
    
     상  하
  상 36   4
  하  7 110
```

```
> prop.table(lpga.qda.ct,1) #행 퍼센트,분류표
    
         상       하
  상 0.90000 0.10000
  하 0.05983 0.94017
```

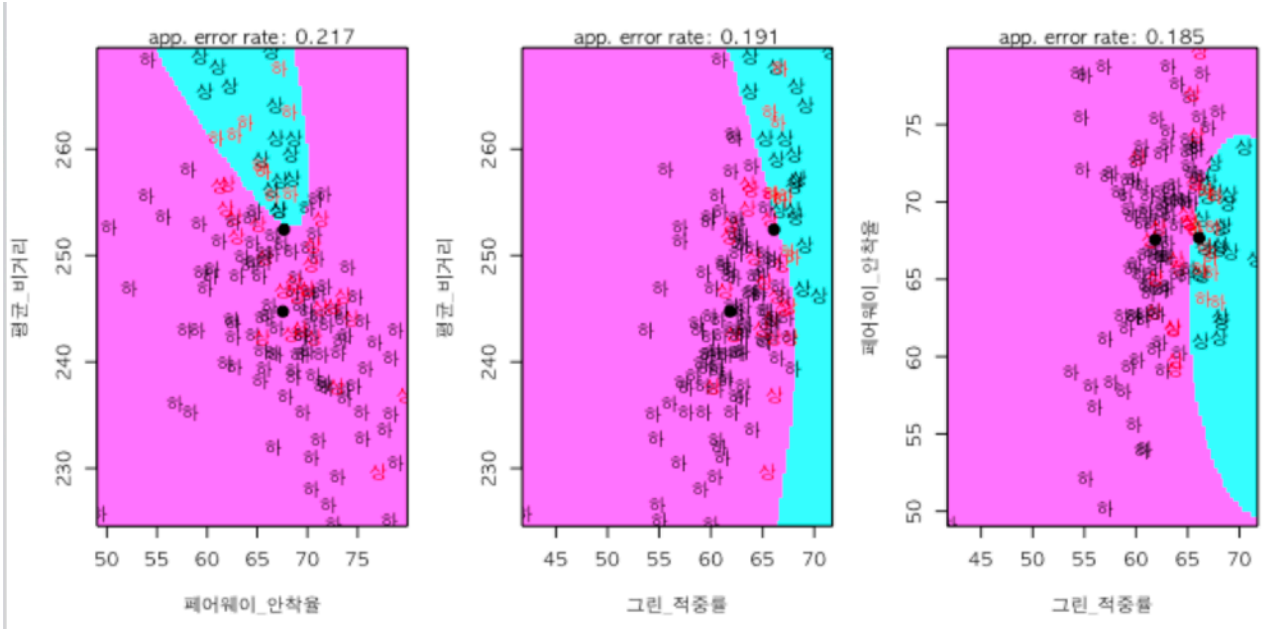
전체 선수 정분류 비율 = (36+110)/157 = 92.9% (선형판별=91.1%)보다 1.8% 높음

```
> sum(diag(prop.table(lpga.qda.ct))) #정분류
[1] 0.9299
```

시각화 (판별변수 쌍체 산점도, 판별규칙, 오분류 개체 표현)

```
library(klaR)
partimat(as.factor(상금그룹)~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수
+샌드_회수+샌드_세이브,data=lpga,method='qda')
```

모든 판별변수(평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브) 쌍으로 산점도에 원 집단을 개체 레이블로 표시하고 (선형) 판별규칙을 나타낸다. 그리고 판별규칙에 의해 오분류 개체(선수)는 붉은 색으로 레이블 한다. 그리고 오분류 비율을 산점도 제목에 출력한다. 다음 산점도는 총 15개($_6C_2$) 산점도 중 일부이다. 판별변수 (비거리, 페어웨이) 2개만으로 도출된 판별규칙은 색을 구별하는 직선이고 이 판별규칙에 의한 오분류 비율은 21.7%(선형=22.9%)이다. (비거리, 그린적중률) 판별규칙은 오분류 비율은 19.1% (선형=20.4%), (페어웨이, 그린)의 오분류는 18.5%(선형=19.7%) -> 2개 변수만으로 판별한다면 (페어웨이, 그린)이 가장 좋다. 그리고 이차판별의 정분류 비율이 선형보다 높다.



오분류 개체 판별

```
lpga.qda.result<-cbind(lpga,lpga.qda.p$class)
lpga.qda.result$분류결과[lpga.qda.result[,10]==lpga.qda.result[,11] &
lpga.qda.result[,10]=='상']<-'OO'
lpga.qda.result$분류결과[lpga.qda.result[,10]==lpga.qda.result[,11] &
lpga.qda.result[,10]=='하']<-'XX'
lpga.qda.result$분류결과[lpga.qda.result[,10]!=lpga.qda.result[,11] &
lpga.qda.result[,10]=='상']<-'OX'
lpga.qda.result$분류결과[lpga.qda.result[,10]!=lpga.qda.result[,11] &
lpga.qda.result[,10]=='하']<-'XO'
table(lpga.qda.result[,12])
```

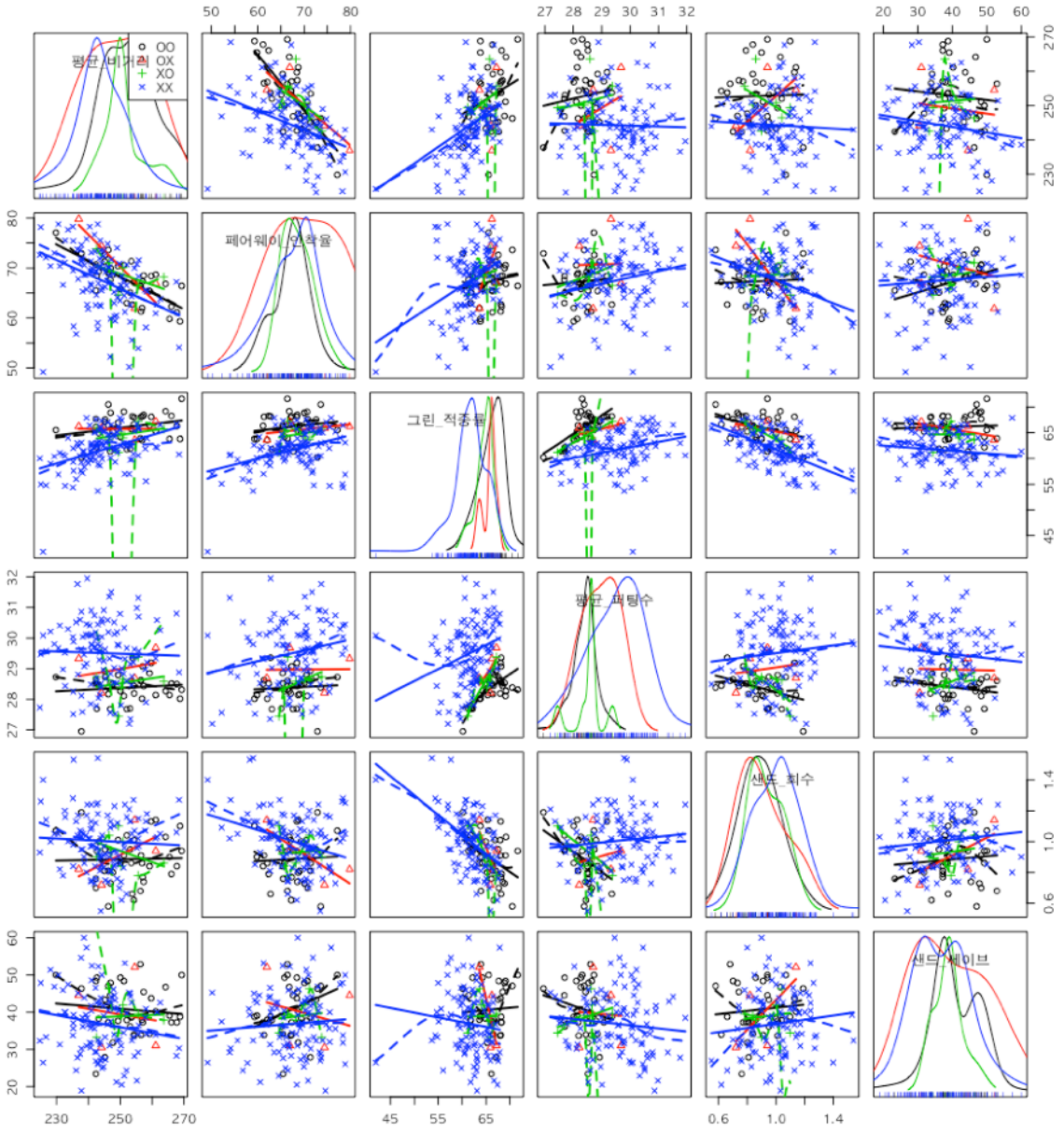
새로 만들어진 '분류결과' 변수는 4개 범주 그룹 변수임

```
> table(lpga.qda.result[,12])

 OO  OX  XO  XX
 36   4   7 110
```

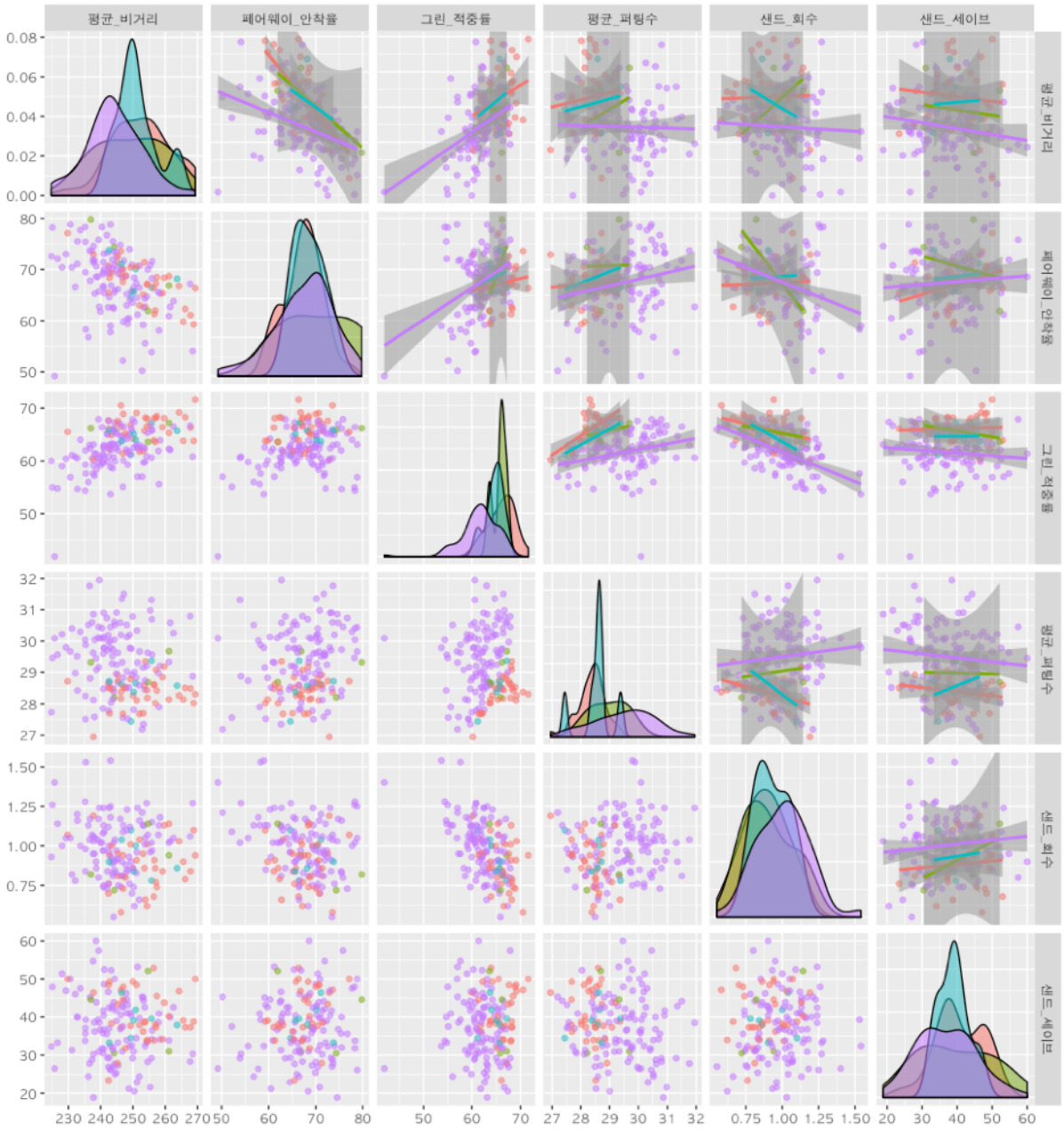
- OO=개체 원집단 '상' - 분류결과 '상', XX=개체 원집단 '하' - 분류결과 '하' - 정분류
- OX=개체 원집단 '상' - 분류결과 '하', OX=개체 원집단 '하' - 분류결과 '상' - 오분류

판별결과 그룹별 산점도

```
scatterplotMatrix(~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수
+샌드_세이브|분류결과,data=lpga.qda.result,col=c(1:4))
```



```
library(GGally)
theme_update(text=element_text(family="AppleGothic"))
ggpairs(lpga.qda.result[,2:7],aes(colour=lpga.qda.result$분류결
과,alpha=0.4),upper=list(continuous='smooth'))
```



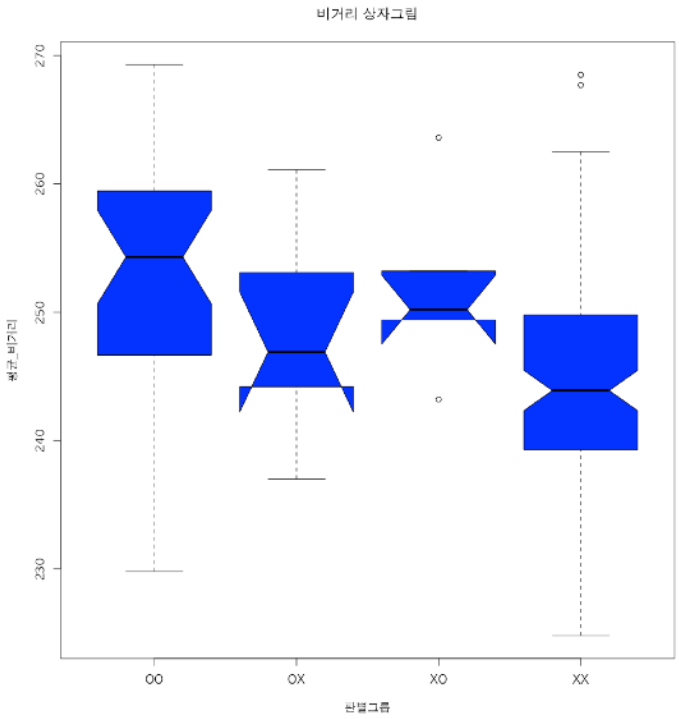
판별결과 그룹별 판별변수 평균, 표준편차

```
library(doBy)
summaryBy(평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브~분류결과,data=lpga.qda.result,FUN=c(mean,sd),na.rm=TRUE)
```

비거리 : 비거리가 큰 선수는 상금 상위 랭킹되므로 평균 252.8 선수는 정분류되지만, 비거리가 다소 낮은 선수는(평균 249.2)는 '하'위 선수로 오분류 되었다. 비거리가 낮은 선수는 하위랭킹 선수이므로 실제 244.3 선수는 '하'위 선수로 정분류 되나, 다소 높은 251.3 부근 선수들의 판별규칙에 의해 상위 선수로 오분류 되었음

```
> summaryBy(평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브~분류결과,data=lpg
  분류결과 평균_비거리.mean 페어웨이_안착율.mean 그린_적중률.mean 평균_퍼팅수.mean 샌드_회수.mean 샌드
1       OO          252.8              67.33          66.14          28.38
2       OX          249.2              70.75          65.73          28.98
3       XO          251.3              68.57          64.67          28.54
4       XX          244.3              67.48          61.66          29.51
```

```
boxplot(평균_비거리~분류결과,data=lpga.qda.result,notch=TRUE,col='blue',
        main="비거리 상자그림", xlab="판별그룹")
```



주성분 분석 활용

```
lpga.pca=prcomp(lpga.qda.result[,2:7],scale.=T) #상관계수행렬 이용
summary(lpga.pca) #변동(누적)설명기여율 출력
lpga.pca$rotation #부하 출력
lpga.qda.pca<-cbind(lpga.qda.result,lpga.pca$x)
```

제1 주성분 : 파온 능력(-), 제2주성분 : 장타능력

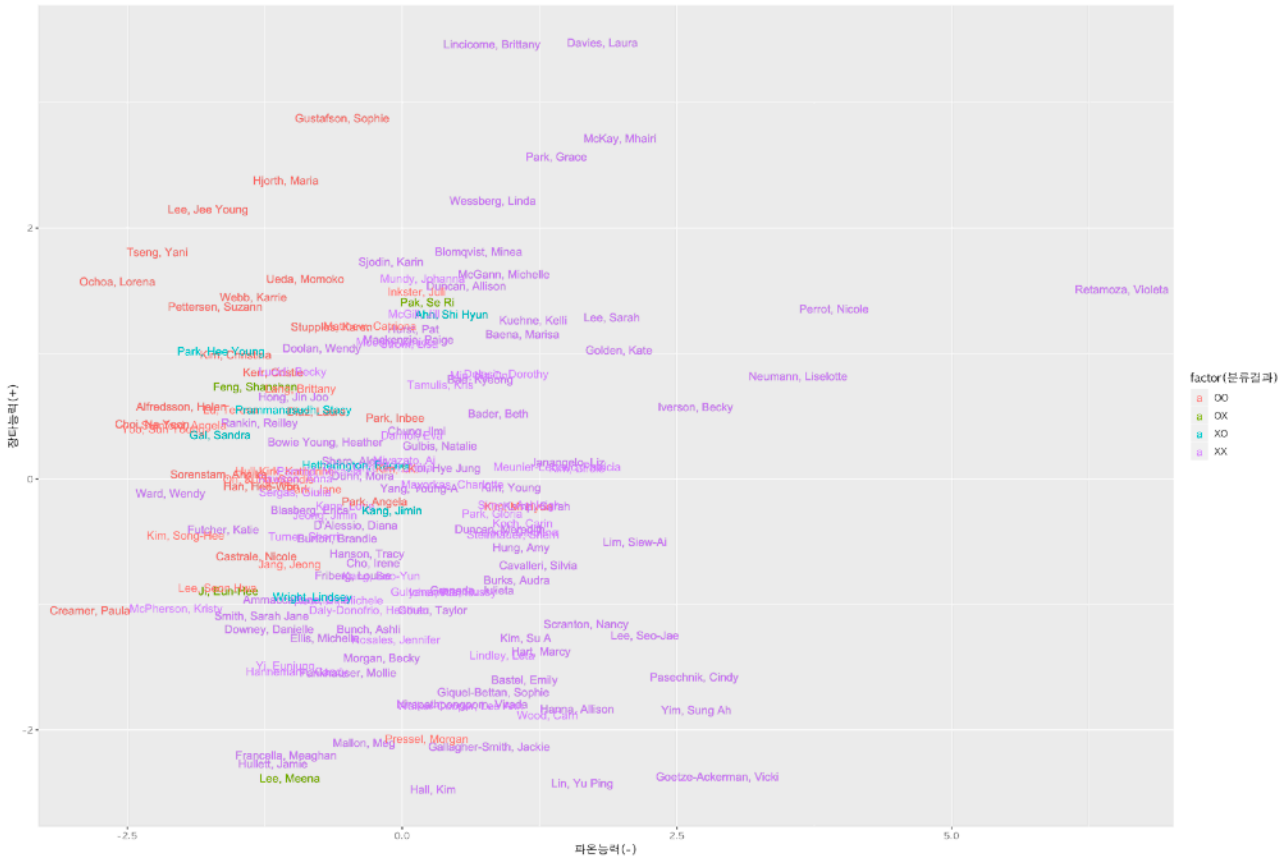
상세 분석 활용 (http://203.247.53.31/Stat_Notes/adv_stat/MDA/MDA_PCA.pdf) 20페이지

```
> summary(lpga.pca) #변동(누적)설명기여율 출력
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6
Standard deviation     1.370 1.239 1.078 0.914 0.6593 0.3921
Proportion of Variance 0.313 0.256 0.194 0.139 0.0724 0.0256
Cumulative Proportion  0.313 0.569 0.763 0.902 0.9744 1.0000
> lpga.pca$rotation #부하 출력
                      PC1       PC2       PC3        PC4      PC5       PC6
평균_비거리      -0.39976  0.57931 -0.15475  0.321000  0.2122 -0.57670
페어웨이_안착율 -0.19523 -0.70654  0.12124 -0.009798  0.5203 -0.42097
그린_적중률      -0.66579 -0.07057 -0.10126  0.277574  0.2082  0.64892
평균_퍼팅수       0.14478 -0.35862 -0.59873  0.582617 -0.3738 -0.11317
샌드_회수         0.58124  0.16983 -0.03577  0.387920  0.6534  0.23363
샌드_세이브       0.00242 -0.05264  0.76899  0.574362 -0.2725 -0.04148
```

```
library(ggplot2)
ggplot(lpga.qda.pca,aes(PC1,PC2,label=골퍼)) +geom_text(check_overlap =
TRUE)+geom_text(aes(colour = factor(분류결과))) +xlab('파온능력(-)')+ylab('장타능
력(+)')
subset(lpga.qda.pca,분류결과=='OX')[,c(1,12)]
```

```
> subset(lpga.lda.pca,분류결과=='OX'
                  골퍼 분류결과
14      Castrale, Nicole       OX
35        Feng, Shanshan       OX
59  Hull-Kirk, Katherine       OX
68           Ji, Eun-Hee       OX
83         Kung, Candie        OX
86           Lee, Meena        OX
95           Lu, Teresa        OX
99     Matthew, Catriona       OX
120          Park, Inbee       OX
```

```
> subset(lpga.qda.pca,분류결과
                  골퍼 분류결과
35  Feng, Shanshan         OX
68      Ji, Eun-Hee        OX
86       Lee, Meena        OX
115      Pak, Se Ri        OX
```



새로운 개체 선수 판별규칙에 의한 사후확률 추정

```
predict(lpga.qda,newdata=data.frame(평균_비거리=260,페어웨이_안착율=70,그린_적중률=65,평균_퍼팅수=28,샌드_회수=1.5,샌드_세이브=40))$posterior
```

```
> predict(lpga.qda,
70,그린_적중률=65,평균_
      상       하
1 0.9709 0.02914
```

비거리 260, 페어웨이 70, …, 샌드세이브=40 인 New 선수는 상위 랭크 가능성은 97.1%(선형=86.1%), 하위랭크 가능성은 2.9%(선형=13.9%)로 판별된다.

모집단 3개 이상인 경우 판별분석

1) ECM 방법

- p_i = 모집단 π_i의 사전확률
- $c(k \mid i)$ = 모집단 π_i의 개체를 π_k로 판별하여 발생하는 비용
- $P(k \mid i)$ = 모집단 π_i의 개체를 π_k로 판별하는 확률

최소화 ECM $= \sum_{i=1}^{g} p_i \left(\sum_{k=1, k\neq i}^{g} P(k \mid i) c(k \mid i) \right) \Leftrightarrow \min \sum_{k=1, k\neq i}^{g} p_i f(k \mid i) c(k \mid i))$

(판별규칙) 만약 $p_k f_k(\underline{x}) \geq p_i f_i(\underline{x})$ for all $i \neq k$ 이면 개체를 모집단 π_k로 판별

2) Fisher의 선형판별식

$$\min_{\underline{\hat{a}}'} \frac{\underline{\hat{a}}'(B = \sum_i^g (\underline{\bar{x}}_i - \underline{\bar{x}})(\underline{\bar{x}}_i - \underline{\bar{x}})')\,\underline{\hat{a}}}{\underline{\hat{a}}'(W = \sum_i^g \sum_i^{n_i} (x_{ij} - \underline{\bar{x}}_i)(x_{ij} - \underline{\bar{x}}_i)')\,\underline{\hat{a}}}$$, B=그룹간 공분산, W=그룹내 공분산

=> $W^{-1}B$ 의 고유값 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_g$, 이에 대응하는 고유벡터가 $\underline{\hat{a}}_1, \underline{\hat{a}}_2, \cdots, \underline{\hat{a}}_g$

k-번째 개체 판별 : $\underline{y}_k = \underline{a}_k' X$, X =데이터 행렬

$\sum_{j=1}^{r} [\underline{\hat{a}}_j'(\underline{x}_0 - \underline{\bar{x}}_k)]^2 \leq \sum_{j=1}^{r} [\underline{\hat{a}}_j'(\underline{x}_0 - \underline{\bar{x}}_i)]^2$ for all $i \neq k$ 을 만족하면 개체 $\underline{x}_0$는 모집

단 π_k로 판별 (r은 고유값 변동기여율 80%인 개수, $r < g$)

3) IRIS 예제 데이터

다변량 데이터셋, 영국 통계학자, 생물학자 Ronald Fisher(1936) n=150 samples

분꽃 3종 (Iris setosa, Iris virginica and Iris versicolor)

4개 판별변수 : Sepal (꽃받침) 넓이, 길이 / Petal 꽃잎 넓이, 길이

```
iris<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/iris.csv')
names(iris)
table(iris$group)
```

```
> names(iris)
[1] "Sepal_Length" "Sepal_Width"  "Petal_Length" "Petal_Width"  "group"
> table(iris$group)

    Setosa Versicolor  Virginica
        50         50         50
```

(1) 등분산 검정

```
library(heplots)
boxM(iris[,1:4],iris$group)
```

유의확률 <0.001로 귀무가설(등분산 가정) 기각되어 이분산 판별식인 이차식 판별이 적절함

```
> boxM(iris[,1:4],iris$group)

        Box's M-test for Homogeneity of Covariance Mat

data:  iris[, 1:4]
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

(2) 피셔 이차판별분석

```
library(MASS)
iris.qda<-qda(group~Sepal_Length+Sepal_Width+Petal_Length+Petal_Width ,
data=iris)
iris.qda.p<-predict(iris.qda) #사후확률($posterior), 판별결과($class)
```

사후확률이 가장 큰 집단(품종)으로 분류됨

```
> head(iris.qda.p$posterior)
          Setosa   Versicolor    Virginica
1  1.000000e+00 5.408818e-23 2.338854e-38
2 1.484542e-165 1.834952e-07 9.999998e-01
3  1.755678e-90 9.973438e-01 2.656243e-03
4 7.824244e-175 1.043025e-09 1.000000e+00
5 4.550670e-111 6.049611e-01 3.950389e-01
6  1.000000e+00 2.633019e-21 2.510567e-34
> head(iris.qda.p$class)
[1] Setosa     Virginica  Versicolor Virginica  Versicolor Setosa
Levels: Setosa Versicolor Virginica
```

(3) 판별결과 정분류 비율

```
iris.qda.ct<-table(iris$group,iris.qda.p$class)
iris.qda.ct
prop.table(iris.qda.ct,1) #행 퍼센트,분류표
sum(diag(prop.table(iris.qda.ct))) #정분류
```

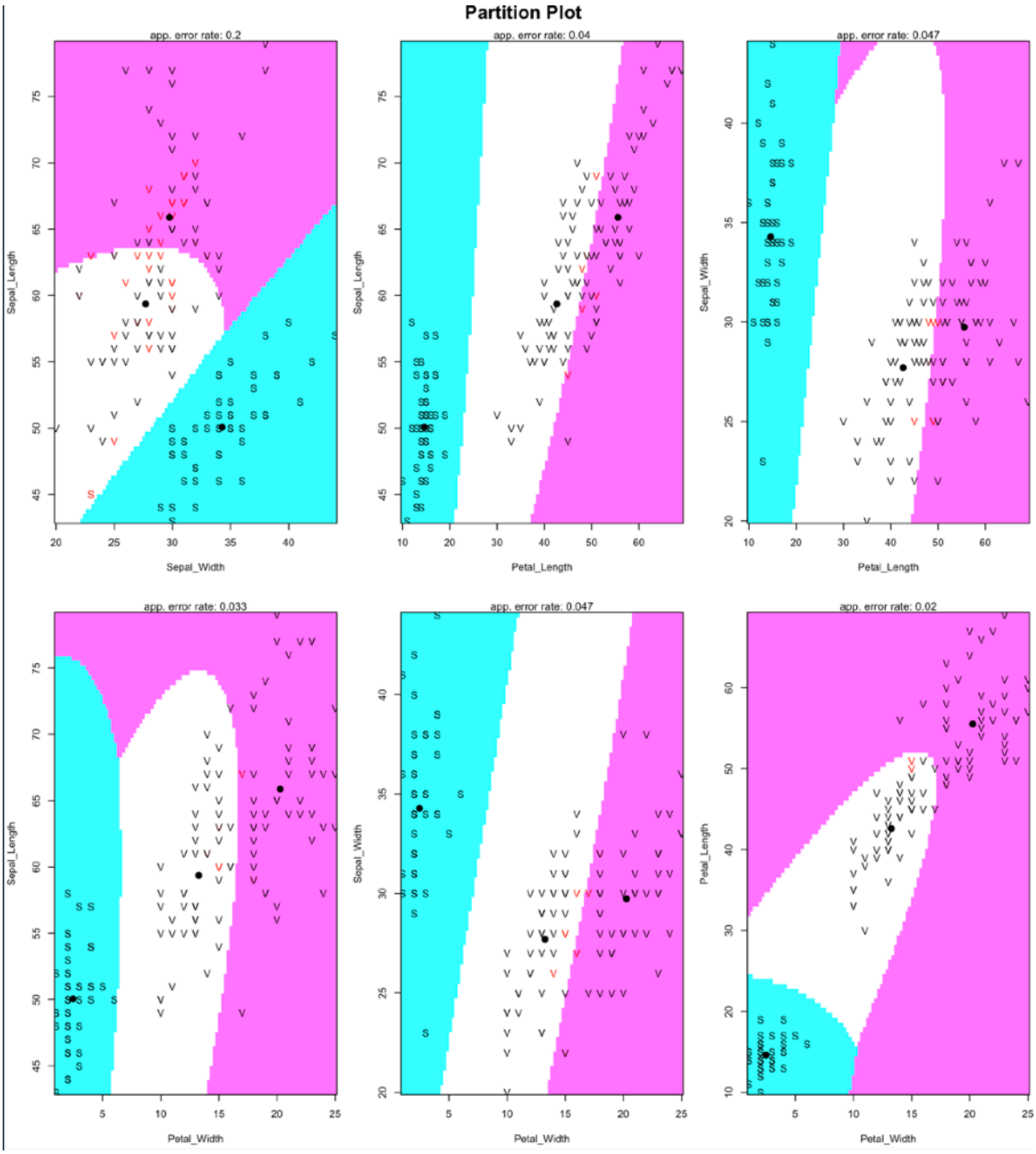
```
> iris.qda.ct

            Setosa Versicolor Virginica
  Setosa        50          0         0
  Versicolor     0         48         2
  Virginica      0          1        49
> prop.table(iris.qda.ct,1) #행 퍼센트,분류표

            Setosa Versicolor Virginica
  Setosa      1.00       0.00      0.00
  Versicolor  0.00       0.96      0.04
  Virginica   0.00       0.02      0.98
> sum(diag(prop.table(iris.qda.ct))) #정분류
[1] 0.98
```

이차 판별 비중은 98%, 매우 높음

(4) 판별결과 시각화 (판별변수 쌍체 산점도 with 판별규칙)

```
library(klaR)
partimat(as.factor(group)~Sepal_Length+Sepal_Width+Petal_Length+Petal_Width,data=iris,method='qda')
```

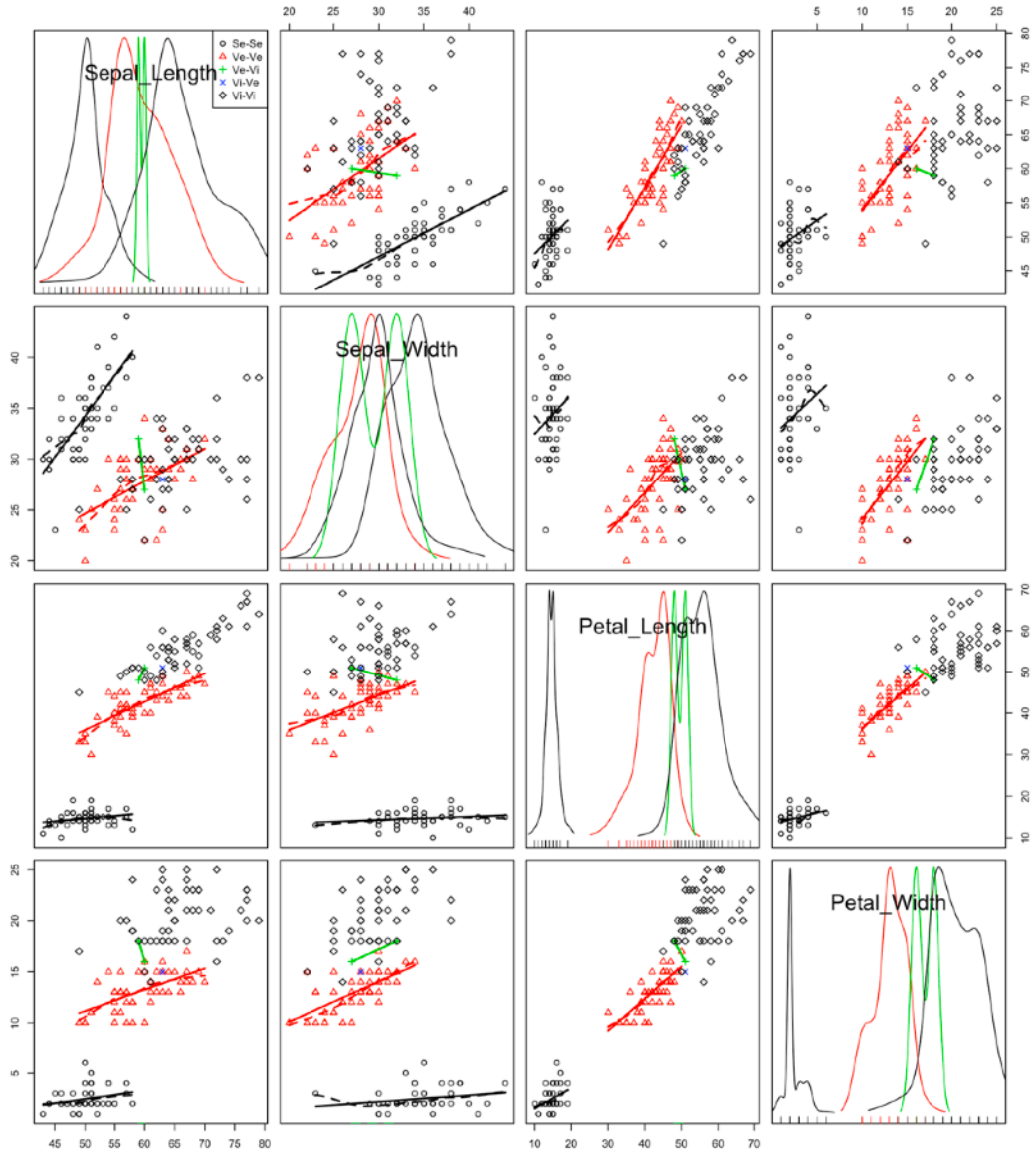
sepal(꽃받침) 넓이, 길이 산점도 - 집단이 3개이고 판별 규칙이 2차식이므로 곡선, 오분류 비율은 20%임.



(5) 판별결과 시각화 2 (판별변수 산점도 행렬)

```
iris.qda.result<-cbind(iris,iris.qda.p$class)
iris.qda.result$classNew<-
paste(substr(iris.qda.result$group,1,2),'-',substr(iris.qda.result[,6],1,2),sep='')
table(iris.qda.result$classNew)
scatterplotMatrix(~Sepal_Length+Sepal_Width+Petal_Length+Petal_Width|
classNew,data=iris.qda.result,col=c(1:4))
```

sepal(꽃받침) 길이는 setosa(o-짧음)와 (virginica(△)+versicolor(✦)-김)군 종을 잘 분류함, 넓이 종의 분류 하지 못함, petal(꽃잎) 길이, 넓이는 setosa(o-길이짧고,넓이좁고), virginica(△), versicolor(✦)길이길고,넓이넓음) 각각 종을 잘 분류함.



Se-Se : Setosa 품종이 Setosa 품종으로 정분류

Ve-Vi : Versicolor 품종이 Virginica 품종으로 오분류

```
library(GGally)
theme_update(text=element_text(family="AppleGothic"))
ggpairs(iris.qda.result[,1:4],aes(colour=iris.qda.result$classNew,alpha=0.4),upper=list(continuous='smooth'))
```

해석은 이전 산점도와 동일함

Moreover, sepal(꽃받침) 길이와 넓이의 직선관계는 비슷(기울기), petal(꽃잎) 길이와 넓이의 직선관계는 종별로 (기울기) 상이.

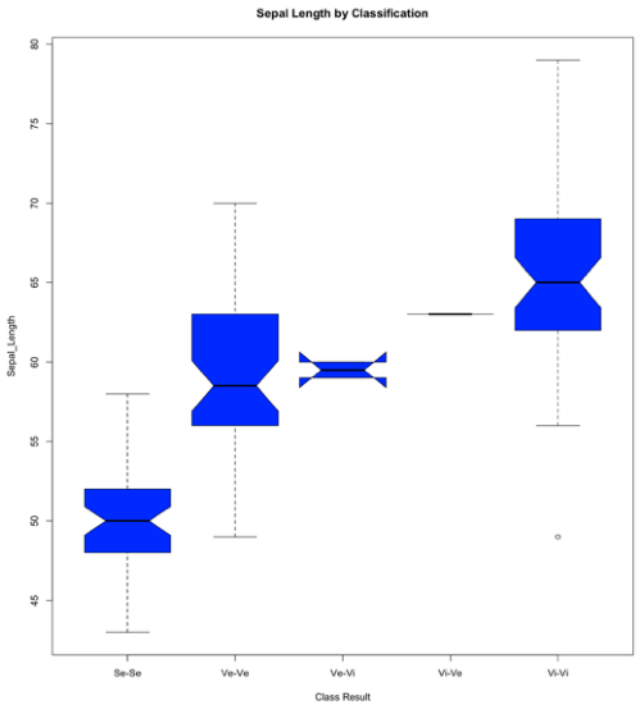
(6) 판별결과 기초통계량

```
library(doBy)
summaryBy(Sepal_Length+Sepal_Width+Petal_Length+Petal_Width~classNew,
data=iris.qda.result,FUN=c(mean,sd),na.rm=TRUE)
```

```
  classNew Sepal_Length.mean Sepal_Width.mean Petal_Length.mean
1    Se-Se          50.06000         34.28000          14.62000
2    Ve-Ve          59.35417         27.62500          42.31250
3    Ve-Vi          59.50000         29.50000          49.50000
4    Vi-Ve          63.00000         28.00000          51.00000
5    Vi-Vi          65.93878         29.77551          55.61224
  Sepal_Length.sd Sepal_Width.sd Petal_Length.sd Petal_Width.sd
1       3.5248969       3.790644        1.736640       1.053856
```

(7) 판별결과 판별변수 산점도

```
names(iris.qda.result)
boxplot(Sepal_Length~classNew,data=iris.qda.result,notch=TRUE,col='blue',
        main="Sepal Length by Classification", xlab="Class Result")
```



(8) 주성분 분석 활용 판별결과 시각화

```
iris.pca=prcomp(iris.qda.result[,1:4]) #공분산행렬 이용
summary(iris.pca) #변동(누적)설명기여율 출력
iris.pca$rotation #부하 출력
iris.qda.pca<-cbind(iris.qda.result,iris.pca$x)
```

주성분 하나로 4개 판별변수의 변동을 92.5% 설명 - 충분함

```
> summary(iris.pca) #변동(누적)설명기여율 출력
Importance of components:
                          PC1     PC2    PC3     PC4
Standard deviation     20.5627 4.92616 2.7966 1.54386
Proportion of Variance  0.9246 0.05307 0.0171 0.00521
Cumulative Proportion   0.9246 0.97769 0.9948 1.00000
> iris.pca$rotation #부하 출력
                    PC1         PC2         PC3         PC4
Sepal_Length  0.36138659 -0.65658877  0.58202985 -0.3154872
Sepal_Width  -0.08452251 -0.73016143 -0.59791083  0.3197231
Petal_Length  0.85667061  0.17337266 -0.07623608  0.4798390
Petal_Width   0.35828920  0.07548102 -0.54583143 -0.7536574
```

```
library(ggplot2)
ggplot(iris.qda.pca,aes(PC1,PC2,label=classNew)) +
  geom_text(check_overlap = TRUE)+
  geom_text(aes(colour = factor(classNew))) +
  xlab('꽃잎길이(-)')+ylab('꽃받침크기(-)')
```

'꽃받침 길이' 성분은 Setosa (가장 적음), Versicolor(중간 크기), Virginica(가장 큼) 3 품종을 잘 분류한다. 그러나 '꽃잎 크기' 성분은 품종 판별 능력이 없음

(8) 새로운 분꽃 분류 집단 추정

```
predict(iris.qda,newdata=data.frame(Sepal_Length=45,Sepal_Width=30,Petal_Length=30,Petal_Width=15))$posterior
```

꽃받침 길이 45, 넓이 30, 꽃잎 길기=30, 넓이=15인 새로운 분꽃이 어떤 종인지 판별하고자 하여 위의 이차 판별식을 활용하여 추정하면 Virginica 58.6%로 가장 크므로 이 품종으로 추정하면 됨

```
        Setosa Versicolor Virginica
1 2.278277e-38  0.4139573 0.5860427
```

단계적 판별분석 방법

판별 분석에서 많은 판별변수(항목)들이 측정되었을 경우 아마 여러분은 의문이 생길 것이다. (1)모든 변수가 판별에 필요한가? (2)어떤 변수가 가장 판별을 잘하는 변수인가? 판별에 적절한 변수만 사용하는 것이 좋으며 판별 능력은 분산 분석의 개념을 이용하여 판단한다.

[선택이유 : Parsimony Rule] 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.

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 F-값이 가장 큰 변수를 선택한다. 공변량은 종속 변수에서 그 변수의 효과를 제외할 때 사용되므로 첫 번째 선택된 변수의 판별 효과를 제외하는 것을 의미한다.

공분산분석: 수학 강의에 대한 새로운 교육방법이 제안되었다. 기존의 교육 방법보다 나은지 알아보기 위하여 각 20명씩 2개의 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 F-값이 가장 큰 변수를 선택한다. 이렇게 변수 선택을 반복한다. 만약 F-값이 가장 큰 것이 유의하지 않으면 (p-값이 유의수준보다 크면) 변수 선택을 멈춘다.

2) Backward 방법

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

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

(2) 같은 방법으로 변수를 하나씩 제거해 간다. 집단의 F-값이 모두 유의하면 (p-값이 유의수준보다 작으면) 제거를 멈춘다.

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) 변수 선택 시 주의점

변수의 수가 15개 이상인 경우 Backward 방법, 15개 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 (X1, X2) 대해 집단간 산점도가 그림(1)과 같다면 변수 X1이 집단을 가장 잘 분류하므로 먼저 선택되고 X2가 그 다음으로 선택된다. 모두 유의하다. 별 문제가 없어 보인다.

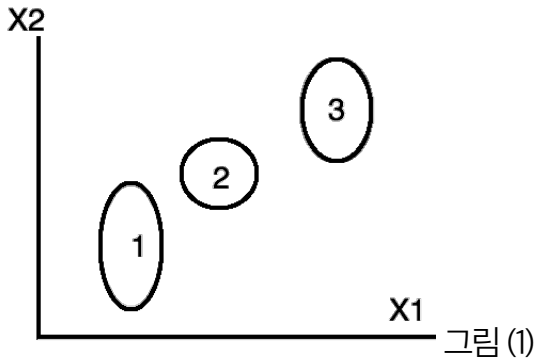


그림 (1)

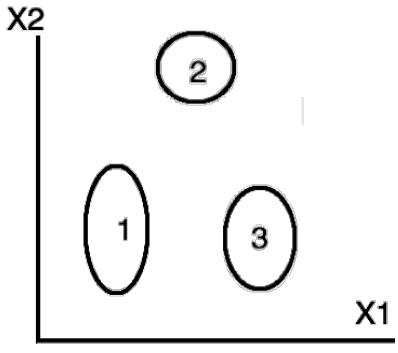


그림 (2)

만약 산점도가 그림(2)와 같다면 실제로는 X1(집단 1,2,3을 모두 분류)가 더 나은 판별을 하지만 변수 선택 방법으로 선택하면 X1(1,3과 2에 집단간 차이가 커서)이 선택될 것이다. 그러므로 변수 선택 방법에 의해 선택된 변수가 판별을 잘한다는 보장은 없다. 그럴지라도 변수 선택 방법은 하나의 좋은 기준을 제시한다.

5) 예제 : LPGA 데이터

등분산 가정이 무너져 6개 판별변수 활용 이차판별규칙 적용

평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브

```
> prop.table(lpga.qda.ct,1) #행 퍼센트,분류표

          상       하
  상 0.90000 0.10000
  하 0.05983 0.94017
```

```
> sum(diag(prop.table(lpga.qda.ct))) #정분류
[1] 0.9299
```

정분류 비율 92.9%

(1) 유의변수 판별

```
library(klaR)
lpga.x<-lpga[,2:7] #판별변수
lpga.y<-lpga[,10]
stepclass(lpga.x,lpga.y,'qda')
```

method='forward', 'backward' 옵션 가능, 디폴트는 'both'

```
method       : qda
final model  : lpga.y ~ 그린_적중률 + 평균_퍼팅수
<environment: 0x7f956ef7fbb8>

correctness rate = 0.8929
```

6개 판별변수 중 2개 변수가 판별에 유의한 변수임. 정분류 비율은 89.29%

(2) 등분산 검정

```
library(heplots); boxM(lpga[,c(2,4)],lpga$상금그룹)
```

등분산 가정 기각 -> 이차판별규칙(함수) 적용

```
> boxM(lpga[,c(2,4)],lpga$상금그룹)

        Box's M-test for Homogeneity of Covariance Ma

data:  lpga[, c(2, 4)]
Chi-Sq (approx.) = 8.6339, df = 3, p-value = 0.03458
```

(3) 2차판별분석 - 등분산 가정 무너짐

```
library(MASS)
lpga.qda2<-qda(상금그룹~그린_적중률+평균_퍼팅수,data=lpga)
lpga.qda.p2<-predict(lpga.qda2) #사후확률(posterior), 판별값(x)
lpga.qda.ct2<-table(lpga$상금그룹,lpga.qda.p2$class)
prop.table(lpga.qda.ct2,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.qda.ct2))) #정분류
```

(그린 적중률, 평균 퍼팅수) 2개 판별변수, 이차판별규칙을 도출하고 정분류 비율을 계산한 결과 91.7%이다. (왜? 앞의 stepclass() 결과와 다르지? stepclass() 함수는 정분류(오분류) 추정을 cross-validation 방법으로 계산하나(사실 이것이 정확한 방법임-그래서 빅데이터에서는 train dataset, test dataset 2분류 하여 test dataset로 오분류 계산) 여기서는 Re-substitution 방법을 적용하였음. 항상 Re-sub 방법이 오분류 비율이 낮음(<-> 정분류 비율이 높음)

```
> prop.table(lpga.qda.ct2,1) #행 퍼센트,분류표

            상          하
  상 0.85000000 0.15000000
  하 0.05982906 0.94017094
> sum(diag(prop.table(lpga.qda.ct2))) #정분류
[1] 0.9171975
```

6개 변수 모두를 사용하여 판별규칙을 만드는 경우 정분류 비율 92.9%에 비해 2개 변수만 사용하여 도출된 판별규칙(함수)의 정분류 비율은 91.7%로 아주 조금 낮다(하 집단은 94%로 동일, 상집단 분류 90%보다 85%로 낮음). 그러나 향후 새로운 선수(개체)에 대한 판별 시 2개 변수(그린 적중률, 평균 퍼팅 수) 값만 측정하면 선수의 집단을 판별할 수 있어 경제성(비용) 측면에서는 2개 판별변수 만으로 추정하는 것이 적절하다. (Parsimony Rule)

새로운 선수 집단 예측

```
predict(lpga.qda2,newdata=data.frame(그린_적중률=65,평균_퍼팅수=28))
$posterior
```

(그린_적중률=65,평균_퍼팅수=28) 새로운 선수를 위에서 얻은 이차판별규칙에 의해 상위랭킹 사후확률이 0.82이므로 이 선수는 상위랭킹 선수가 될 것으로 추정된다.

```
> predict(lpga.qda2,newdata=data.frame(그린_적중률=65,평균_퍼팅수=28))$posterior
         상          하
1 0.8187294 0.1812706
```

이하 분류결과 판별변수 기초통계량 구하기(summaryBy()함수) 시각화 등은 이전 이차 판별분석 결과를 참고하시오. (페이지 38-40)

로지스틱 회귀모형

1) 모형 (이진형 목표변수)

판별변수 벡터 $\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ 가 설명변수, 모집단 그룹이 종속변수인 일반화 회귀모형

$y_i = 0, 1$, 모집단1={성공}={1}, 모집단2={실패}={0}, $\pi_i = P(Y_i = ' Success')$

설명변수(예측변수) 개수 p개인 회귀모형

$$y_i = a + b_1 X_{1i} + b_2 X_{2i} + \cdots b_p X_{pi} + e_i$$

$$E(Y \mid \underline{x}) = X\underline{b} + \underline{e}, (\underline{e} \sim N(\underline{0}, \sigma^2 I))$$

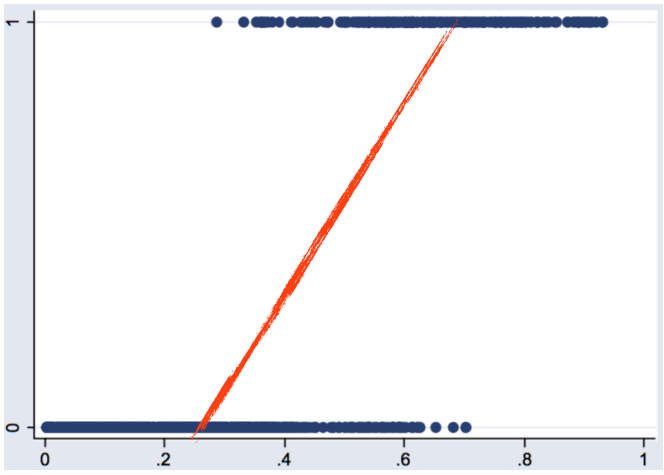
가정 : $e_i \sim N(0, \sigma^2) \Rightarrow y_i \sim \sim N(X\underline{b}, \sigma^2)$

종속변수 y_i가 측정형이 아니라 (성공, 실패)의 베르누이 분포를 따른다면, $E(Y \mid \underline{x}) = p$ 이다.

만약 Y_i= 성공확률이 (π_i-추정)인 i-번째 그룹(총 관측 개수 n_i)에서 관심사건 발생건수라 정의하자.
$Y_i \sim B(n_i, \pi_i)$을 따른다. 그러므로 $E(Y_i) = n_i \pi_i$이다.

2) OLS 추정의 문제 (회귀직선)

회사 내 남자비율이(예측변수, X) 높을수록 보수적 회사일 (1=보수, 0=진보) 가능성이 높다는 가설을 검정하기 위하여 수집한 데이터의 산점도이다. 위의 점은 보수 기업, 아래 점들은 진보 기업이다. X-축은 남자 직원 비율이 된다. 붉은 직선은 OLS(최소자승추정법) 추정된 회귀직선이다. 더 이상 OLS(최소자승추정법)은 MVUE(최소분산불편추정량)을 보장하지 못한다.



3) 회귀계수 추정

연결함수 Link function

연결함수 $g(\mu) = X\underline{b}$ where μ = 종속변수 평균 (여기서는 $\mu = p$)

$g(\mu)$는 종속변수가 $(-\infty, \infty)$의 값을 갖도록 변환하는 함수

3) logit 회귀모형

오즈(odds) : 성공확률을 실패확률로 나눈 값 $odds_i = \dfrac{\pi_i}{1-\pi_i}$

- 두 정수로 표현 4:1, 4-1, 4/1 (4 to 1로 읽음, 성공확률이 0.8) ={성공가능성이 4배 높다. 즉 5번 시행하면 4번 성공함}
- 오즈 On(성공 확률 높음), against(실패 확률 높음), even(확률동일)

오즈는 $(0, \infty)$를 가지므로 로짓변환하여 $(-\infty, \infty)$ 값을 가질 수 있으므로 일반 회귀모형 적합 가능

로짓 logit 혹은 log-odds : $logit(\pi_i) = ln(\dfrac{\pi_i}{1-\pi_i}) = X\underline{b} + \underline{e},\ \underline{e} \sim MN(\underline{0}, \Sigma)$

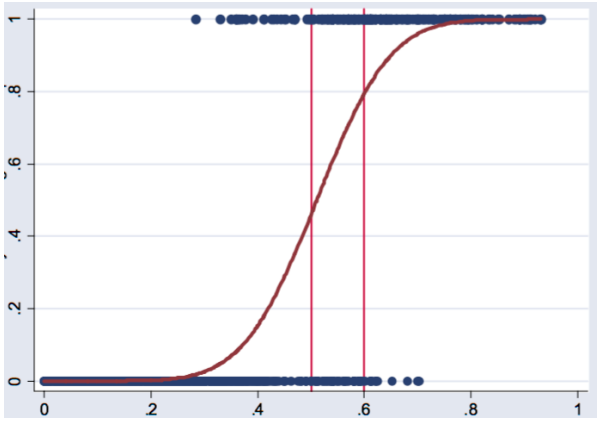
$$\pi_i = \frac{exp(x_i'\underline{b})}{1 + exp(x_i'\underline{b})}$$

회귀계수 추정 : MLE 추정

$\Pr\{Y_i = y_i\} = \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$ 이므로 우도함수는

$\log L(\boldsymbol{\beta}) = \sum\{y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i)\}$ 이다. 그러므로 MLE 추정량은

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}, \quad z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(n_i - \hat{\mu}_i)} n_i, \quad w_{ii} = \hat{\mu}_i(n_i - \hat{\mu}_i)/n_i.$$



회귀계수 $\hat{b}_k$의 의미

- 설명변수 X_k는 개체 성공 확률에 양의 영향을 미친다. (예) 남자직원 비율이 높은 기업은 보수 성향이 높아진다.
- 설명변수 X_k가 한 단위 증가할 때 오즈비는 $e^{\hat{b}_k}$ 배만큼 **곱**으로 증가한다.

4) 오즈비(Odds Ratio)와 상대위험도(Relative risk)

	미취업	취업
토익 750 이상	$5(\pi_1)$	$15(1-\pi_1)$
점수 없음	$10(\pi_2)$	$10(1-\pi_2)$

(1) 상대 위험도 relative risk $\frac{\pi_1}{\pi_2}$ = (10/20)/(5/20)=2

(토익 750 이상) 위험도(미취업비율) = 5/20(=25%)

(점수없음) 위험도(미취업비율) = 10/20(=50%)

- 토익점수 없는 학생은 750 이상인 학생에 비해 미취업 가능성이 2배 높다. (해석이 명확하다)
- 모집단 전체 수를 알고 있는 경우에는 상대 위험도 사용 (졸업 예정자가 40명인 학과 조사연구, 코호트 연구)
- Cohort 연구 : 특정 모집단을 정의하고 연구목적(예: 흡연=원인과 폐암의 관계)을 일정 기간 추정 후 사건(폐암)의 발생여부 관찰하는 연구방법

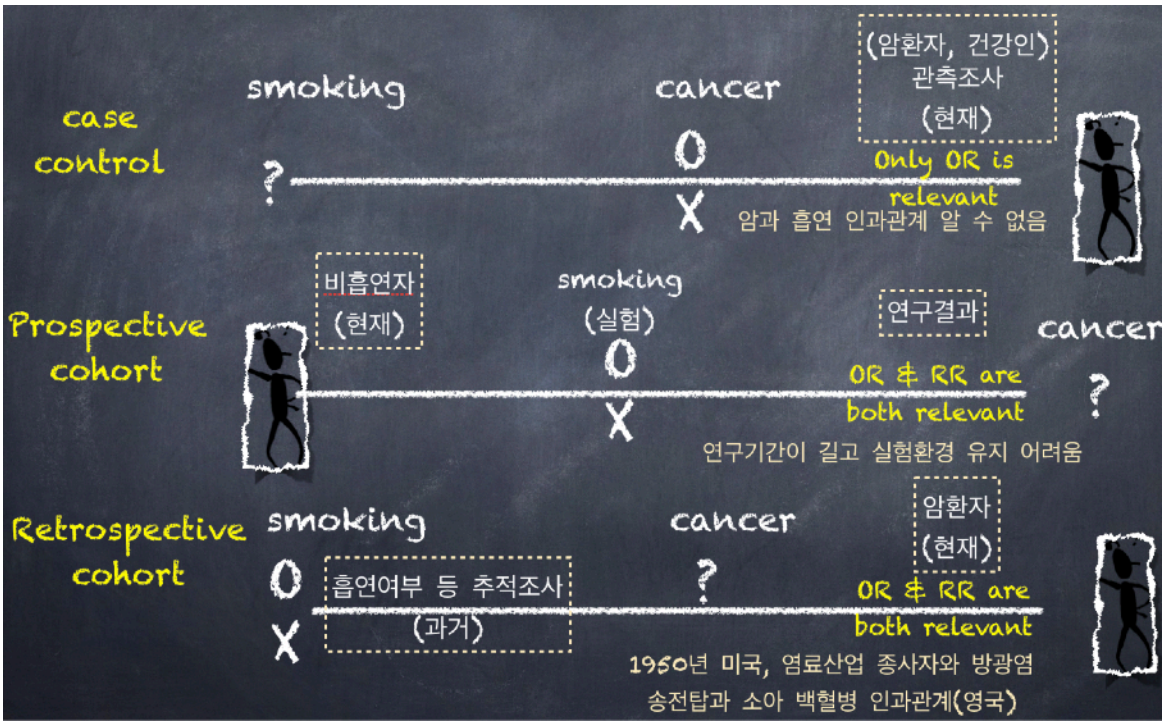
(2) 오즈비 odds ratio =(10/10)/(5/15)=3

(토익 750 이상) 오즈 = 5/15=1/3 (토익 750인 학생의 취업성공율은 미취업비율에 비해 3배 높다)

(점수 없음) 오즈 =1 (취업성공율과 미취업비율은 동일하다)

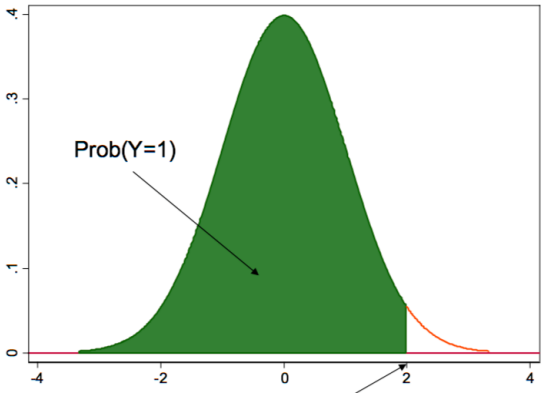
- (해석) 취업에 성공한 사람은 미취업자에 비해(**사건 중심 해석**) 토익점수가 750일 가능성 3배 높음
- 완전한 상대위험도 개념은 아님, 모집단 전체 수를 모르는 환자-대조 실험연구에서 상대위험도 개념으로 사용, 상대 위험도 과다 추정 (over-estimate), 건강(취업) 대비 미취업(환자) 수가 적어지면 오즈비와 상대위험도는 근사한다.

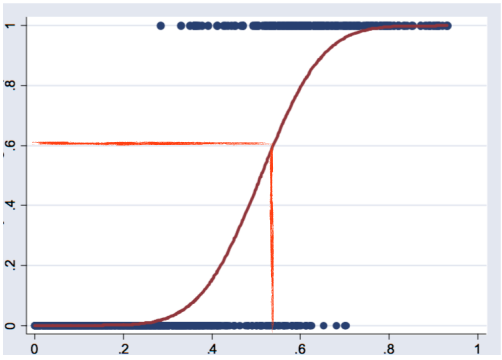
[사례연구 case control] 이미 발생한 사건(폐암 여부) 발생 여부에 의해 control(통제, 비흡연) 집단과 experiment(실험, 흡연) 집단으로 나누고 원인(흡연)과 사건(폐암)의 연관성을 조사하는 것



5) probit 변환 $\Phi^{-1}(\mu) = \Phi^{-1}(p) = X\underline{b}$

Φ 는 누적표준정규분포함수이며 Φ^{-1} 는 정규분포의 확률변수 값이므로 $(-\infty, \infty)$ 을 가짐





(활용) 보수성향 60%이 기업의 남자 사원 비율은 55%이다. 즉 사원비율이 55%이면 보수성향은 60%이다. 주로 의학분야 선형모형 적용 시 사용함

로지스틱 판별분석

1) 개념

모집단이 2개인 경우

- 모집단1={성공, π_1}, 모집단2={실패, π_2}, 성공확률 $p = P(y_i \subseteq \pi_1)$
- 판별변수 = "설명변수"
- 모형 $P(Y_i = \left\{ sucess \right\}) = \dfrac{\exp(X\underline{b})}{1+\exp(X\underline{b})}$
- 추정 결과 : 사후확률 $\hat{p} = P(y_i \subseteq \pi_1 \,|\, data)$이 0.5보다 크면 {성공}, 작으면 {실패} 집단으로 분류한다. (일반적 규칙)
- ROC 분석을 실시하여 오분류를 가장 적게 하는 사후확률을 성공, 실패 나누는 기준값으로 사용할 수 있다.

모집단이 3개 이상인 경우

- 모집단 기호 : $g = 1, 2, \ldots, G$, $logit(P(Y \le g)) = log\dfrac{P(Y \le g)}{P(Y > g)} = -X\underline{b}$
- 모형 $P(Y_i \le g) = \dfrac{\exp(X\underline{b})}{1+\exp(X\underline{b})}, g = 1, 2, \ldots, G$
- 절편이 하나가 아니라 ($a_1, a_2, \ldots, a_{G-1}$)로 모집단 개수보다 1개 적음
- 집단변수는 순서형이 해석이 용이하다. (예: 5점 리커트 척도, 알파벳 학점)

Fisher 방법에 비해 장점

- 회귀계수의 부호에 의해 해당 판별변수가 성공 집단에 속할 확률에 미치는 영향 정도를 평가할 수 있음
- 성공 집단에 속할 사후확률(0~1사이 실수)을 추정하여 이를 활용하여 개체의 순위를 결정할 수 있음
- ROC 분석을 통하여 오분류 최소화 하는 "성공/실패" 집단 판별하는 기준값을 설정할 수 있음

2) 로짓분석 - 사례연구 : LPGA 데이터 (상금 랭킹 UP, DN 이진형)

종속변수 : 상위랭킹 여부(상금순위 40위 이내='상', 그외 = '하'

판별변수(독립변수) : 6개-(비거리, 페어웨이 안착율, 샌드(벙커) 세이버, 그린 적중률, 버팅 수)

```
lpga<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/
lpga2008.csv', fileEncoding = "utf-8")
lpga$ranking[rank(-lpga$상금)<=40]<-'UP'
lpga$ranking[rank(-lpga$상금)>40]<-'DN'
```

```
> head(lpga)
                골퍼 평균_비거리 페어웨이_안착율 그린_적중률 평균_퍼팅수
1     Ahn, Shi Hyun       249.4           64.6        61.2       27.44
2 Alfredsson, Helen       253.8           62.7        68.2       29.36
3 Ammaccapane, Dina       246.3           70.2        64.6       30.20
4       Bader, Beth       249.1           64.1        61.2       29.78
5       Bae, Kyeong       244.0           62.4        60.7       28.38
6     Baena, Marisa       254.2           64.7        60.9       29.21
  샌드_회수 샌드_세이브  상금 참가_라운드수 ranking
1      1.10        34.5  6063            50      DN
2      0.66        38.8 19343            74      UP
3      0.74        40.5  1873            50      DN
4      1.12        41.1  1212            65      DN
5      1.02        43.9  2555            65      DN
6      1.27        33.3  2282            52      DN
```

(1) 로짓분석(추정)

```
lpga.logit<-glm(factor(ranking)~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼
팅수+샌드_회수+샌드_세이브, data=lpga,family=binomial(link="logit"))
summary(lpga.logit); names(lpga.logit)
```

로짓분석 모형 추정결과의 요약 리스트

- 성공집단은 알파벳 순위가 뒤인 집단 즉, UP을 성공집단으로 설정한다.
- 그린적중률(+높을수록), 평균 퍼팅수(-작을수록), 샌드세이브(+높을수록)가 성공집단(UP) 속할 확률이 올라간다(유의함).
- (추정) 회귀계수 : lpga.logit$coefficients, 오즈비. exp(lpga.logit$coefficients)

```
> summary(lpga.logit)

Call:
glm(formula = factor(ranking) ~ 평균_비거리 + 페어웨이_안착율
    그린_적중률 + 평균_퍼팅수 + 샌드_회수 + 샌드_세이브,
    family = binomial(link = "logit"), data = lpga)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.17933  -0.06234  -0.00253   0.00012   1.96787

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     24.304511  29.588932   0.821  0.41142
평균_비거리     -0.006224   0.087091  -0.071  0.94303
페어웨이_안착율 -0.102716   0.154007  -0.667  0.50480
그린_적중률      2.147288   0.537701   3.993 6.51e-05 ***
평균_퍼팅수     -5.777019   1.432128  -4.034 5.49e-05 ***
샌드_회수        4.734951   3.559527   1.330  0.18345
샌드_세이브      0.165326   0.062773   2.634  0.00845 **
```

분석결과에 저장된 통계량 리스트

```
> names(lpga.logit)
 [1] "coefficients"      "residuals"         "fitted.values"
 [4] "effects"           "R"                 "rank"
 [7] "qr"                "family"            "linear.predictors"
[10] "deviance"          "aic"               "null.deviance"
[13] "iter"              "weights"           "prior.weights"
[16] "df.residual"       "df.null"           "y"
[19] "converged"         "boundary"          "model"
[22] "call"              "formula"           "terms"
[25] "data"              "offset"            "control"
[28] "method"            "contrasts"         "xlevels"
```

```
> exp(lpga.logit$coefficients)
    (Intercept)      평균_비거리 페어웨이_안착율      그린_적중률
   3.591825e+10     9.937954e-01     9.023828e-01     8.561606e+00
    평균_퍼팅수        샌드_회수       샌드_세이브
   3.097936e-03     1.138579e+02     1.179778e+00
```

(2) 유의변수 선택

```
stepAIC(lpga.logit,direction='both')
```

- step() 함수 사용가능, both 대신 forward, backward 사용가능

최종 단계에 선택된 변수에 의한 추정모형이 출력된다.

```
Coefficients:
(Intercept)   그린_적중률   평균_퍼팅수      샌드_회수   샌드_세이브
    22.8444        2.1235       -5.9781        5.0928        0.1622
```

(3) 유의변수 모형추정

```
lpga.logit0<-glm(factor(ranking)~그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,
data=lpga,family=binomial(link="logit"))
summary(lpga.logit0)
```

AIC (Akai keInformation Criterion) $AIC = 2(p - ln(Likihood))$(충분히 작아지는 변수개수) 기준 변수선택이므로 유의하지 않은 변수(샌드회수)가 출력된다.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 22.84444   19.29310   1.184  0.23638
그린_적중률  2.12349    0.53490   3.970 7.19e-05 ***
평균_퍼팅수 -5.97809    1.49573  -3.997 6.42e-05 ***
샌드_회수    5.09278    3.56044   1.430  0.15261
샌드_세이브  0.16223    0.06231   2.604  0.00923 **
```

```
> lpga.logit0$formula
factor(ranking) ~ 그린_적중률 + 평균_퍼팅수 + 샌드_회수 +
    샌드_세이브
```

(4) 회귀계수 해석 (유의변수)

그린적중율이 높을수록, 퍼팅수가 작을수록, 샌드 세이브 성공 높을수록 상위랭커일 가능성 높다.

```
lpga.logit0<-glm(factor(ranking)~그린_적중률+평균_퍼팅수+샌드_세이브,
                 data=lpga,family=binomial(link="logit"))
summary(lpga.logit0)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 35.75090   17.54356   2.038  0.04157 *
그린_적중률  1.82922    0.44161   4.142 3.44e-05 ***
평균_퍼팅수 -5.62652    1.41580  -3.974 7.06e-05 ***
샌드_세이브  0.17595    0.06037   2.914  0.00356 **
```

- 그린적중률 오즈 6.23, 평균 퍼팅수 0.0036(역수 - 277), 샌드세이브 1.19 (해석) 그린적중율이 높은 선수는 낮은 선수에 비해 성공(상위랭킹) 확률이 6.23배 높다. 평균퍼팅 수가 작은 집단(역수 1 이하이므로)은 많은 집단에 비해 성공 확률은 277배 높다.
- 결론적으로 3개 유의 변수 중 평균 퍼팅수 변인의 (성공=상위랭킹) 영향이 가장 높다.
- **[오즈비 해석은 판별변수의 단위가 다른 경우에는 적절하지 않음]** (그린적중률 오즈)/(샌드세이브 오즈)=6.23/1.19=5.23(배) - (해석) 상위랭킹 속할 가능성은 그린 적중율이 높은 선수가 샌드세이브 높은 선수에 비해 5.23배 높다. [두 변수의 단위는 동일하므로 가능]

```
> exp(lpga.logit0$coefficients)
 (Intercept)   그린_적중률   평균_퍼팅수   샌드_세이브
3.360613e+15 6.229045e+00 3.601094e-03 1.192379e+00
```

(5) 다중공선성 multicollinearity 진단

회귀모형에서 설명변수 간 높은 상관관계로 회귀계수의 추정분산이 커져 추정 회귀계수의 부호까지 바뀌는 현상을 다중공선성[**강의노트 가기**] 이라 한다.

진단

기준값 = 4, 그린적중률, 평균_퍼팅수의 높은 상관관계로 다중공선성 문제 발생 진단

```
> vif(lpga.logit0)
그린_적중률 평균_퍼팅수 샌드_세이브
   6.980928    6.301156    1.845176
```

해결

다중공선성 문제 변수 제거 -> 평균 퍼팅수가 더 유의하나 샌드세이브와 상관관계가 낮고 해석이 용이한 그린_적중율을 남기고 평균 퍼팅수 제거함

(6) 최종 회귀모형 (AIC 기준 stepclass 판별변수 선택)

```
lpga.logit0<-glm(factor(ranking)~그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,
data=lpga,family=binomial(link="logit"))
summary(lpga.logit0)
exp(lpga.logit0$coefficients)
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 22.84444   19.29310   1.184  0.23638
그린_적중률  2.12349    0.53490   3.970 7.19e-05 ***
평균_퍼팅수 -5.97809    1.49573  -3.997 6.42e-05 ***
샌드_회수    5.09278    3.56044   1.430  0.15261
샌드_세이브  0.16223    0.06231   2.604  0.00923 **
```

그린적중율, 샌드 세이브 오즈비 = 8.36/1.17=7.14(배) - 그린 적중율 높은 선수가 샌드 세이브 높은 선수에 비해 성공(상위랭킹) 확률이 7배 이상 높다. (더 큰 영향)

```
> exp(lpga.logit0$coefficients)
 (Intercept)   그린_적중률   평균_퍼팅수     샌드_회수   샌드_세이브
8.340888e+09 8.360254e+00 2.533666e-03 1.628414e+02 1.176133e+00
```

그러나 샌드_횟수(1자리)와 그린 적중률(2자리)의 단위가 다르므로 오즈비 해석은 적절하지 않음 - 만약 단위가 상이한 오즈비 해석이 필요하다면 표준화 회귀계수(판별변수 표준화 후 로짓분석)를 추정하여 오즈비를 해석한다.

(6) 사후확률 추정결과

추정결과는 $fitted.values 저장되어 있음

```
head(lpga.logit0$fitted.values)
#head(predict(lpga.logit0,type='response')) #동일 사후추정 결과 얻음
```

```
> head(predict(lpga.logit0,type='response'))
           1            2            3            4            5
4.905721e-01 8.588236e-01 3.801914e-05 2.616021e-06 3.679464e-03
> head(lpga.logit0$fitted.values)
           1            2            3            4            5
4.905721e-01 8.588236e-01 3.801914e-05 2.616021e-06 3.679464e-03
```

정분류 결과표 사후확률 0.5기준

```
lpga$class0<-ifelse(lpga.logit0$fitted.values>0.5,'UP','DN')
lpga.logit.ct0<-table(lpga$ranking,lpga$class0)
lpga.logit.ct0
prop.table(lpga.logit.ct0,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.logit.ct0))) #정분류
```

4개 판별변수에 의한 로짓판별 결과 정분류 비율은 93.6%임 (피셔 6개 판별변수 활용 이차판별 결과는 92.9%, 페이지 25 참고) 4개만 사용한 로짓판별 규칙의 정분류 비율이 높으므로 가장 적절한 방법임

```
> lpga.logit.ct

     DN  UP
  DN 112   5
  UP   5  35
> prop.table(lpga.logit.ct,1) #행 퍼센트,분류표

            DN         UP
  DN 0.95726496 0.04273504
  UP 0.12500000 0.87500000
> sum(diag(prop.table(lpga.logit.ct))) #정분류
[1] 0.9363057
```

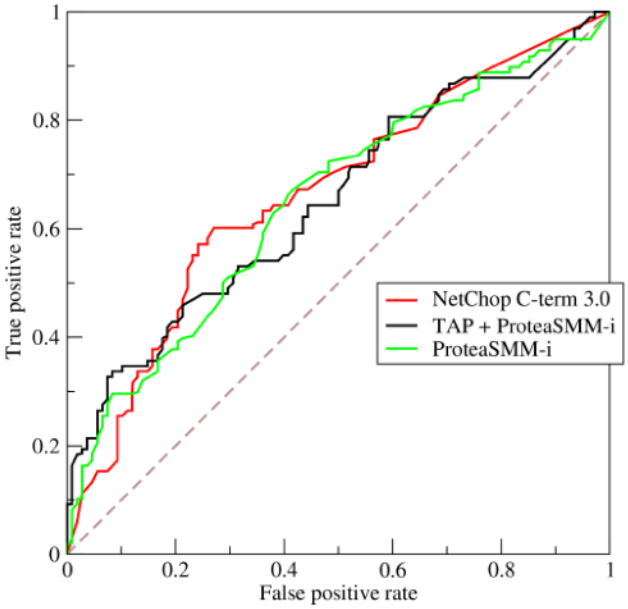
6개 모두 판별변수 사용

```
lpga.logit<-glm(factor(ranking)~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브, data=lpga, family = binomial(link="logit"))
lpga$class<-ifelse(lpga.logit0$fitted.values>0.5,'UP','DN')
lpga.logit.ct<-table(lpga$ranking,lpga$class)
prop.table(lpga.logit.ct0,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.logit.ct))) #정분류
```

4개 유의 판별변수 활용한 결과와 정분류 비율 동일 - Parsimony Rule 에 의해 유의 판별변수 선택하는 것이 적절함

```
> prop.table(lpga.logit.ct,1) #행 퍼센트,분류표

            DN         UP
  DN 0.95726496 0.04273504
  UP 0.12500000 0.87500000
> sum(diag(prop.table(lpga.logit.ct))) #정분류
[1] 0.9363057
```

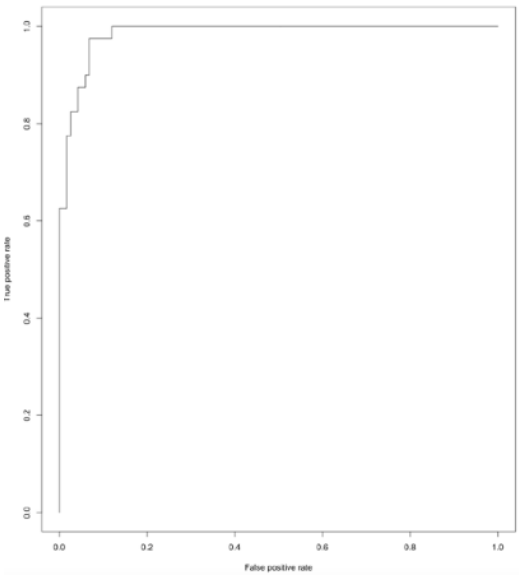
(7) 판별결과 평가 ROC 커브



- True Positive Rate = Sensitivity (민감도) 양성을 양성으로 진단한 정분류
- False Positive Rate = 1- Specificity(특이도)
- 특이도 : 음성을 음성으로 진단한 정분류
- AUC : Area Under Curve ROC 곡선 아래 면적, 회색부분 1에 가까울수록 완벽한 판별
- ROC : Receiver operating characteristic 판별규칙의 기준 확률 0~1 사이 값에서 TPR, FPR
- 왼쪽 ROC 는 3개 판별(진단)규칙의 ROC 곡선이다. (출처 위키피디아)

```
lpga.logit0<-glm(factor(ranking)~그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브,
data=lpga,family=binomial(link="logit"))
library(ROCR)
p<-predict(lpga.logit0,newdata=lpga,type='response')
pr<-prediction(p,lpga$ranking)
prf<-performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf); auc<-performance(pr,measure = "auc"); auc@y.values[[1]]
```

만약 모형 설정 train data, 모형 검증 데이터 test가 있다면 newdata=test를 하면 된다. 4개 판별변수, 로짓 판별규칙의 정확성은 매우 높다.



```
> auc@y.values[[1]]
[1] 0.9844017
```

excellent = 0.9~1
good = 0.8~0.9
fair = 0.7~0.8
poor = 0.6~0.7
fail = 0.5~0.6

(8) 판별결과 판별 그룹별 기초통계량 구하기

```
lpga$classResult<-paste(lpga$ranking,'-',lpga$class0,sep='')
table(lpga$classResult)
library(doBy)
summaryBy(그린_적중률+평균_퍼팅수+샌드_회수+샌드_세이브
~classResult,data=lpga,FUN=c(mean,sd),na.rm=TRUE)
```

```
> table(lpga$classResult)

DN-DN DN-UP UP-DN UP-UP
  112     5     5    35
```

그린적중률 : 상위랭킹이 정분류 선수 평균 그린 적중률 66.4, 하위랭킹 정분류 선수 평균 61.70이므로 하위랭킹의 선수 중 그린 적중률이 높은 선수들(평균 65.2) 상위랭킹으로 오분류 되었고, 상위랭킹 선수 중 그린 적중률이 낮은(64.1) 선수들의 하위 랭킹으로 오분류되었다.평균 퍼팅 수가 작을수록 상위랭킹, 그러므로 상위랭킹 중 퍼팅 수가 많은 선수들은 하위랭킹으로 오분류된다.

```
  classResult 그린_적중률.mean 평균_퍼팅수.mean 샌드_회수.mean
1       DN-DN          61.69286          29.48964       1.0057143
2       DN-UP          65.18000          28.62000       0.8620000
3       UP-DN          64.12000          28.59800       0.8660000
4       UP-UP          66.38571          28.41743       0.8902857
  샌드_세이브.mean 그린_적중률.sd 평균_퍼팅수.sd 샌드_회수.sd 샌드_세이브.sd
1         36.90000       3.789143       1.0477706   0.17935678
2         40.44000       1.886001       0.4957822   0.05761944
3         41.86000       2.558711       0.9159258   0.11696153
4         40.41143       2.484084       0.4516333   0.15778775
```

(9) 새로운 선수 예측

(평균_비거리=260,페어웨이_안착율=70, 그린_적중률=65,평균_퍼팅수=28,샌드_회수=1.5,샌드_세이브=40) - 피셔이차판별규칙 결과 상위랭킹 선수일 확률은 97.1%(페이지 31 참고)

로짓판별함수는 (평균_비거리, 페어웨이_안착율)은 유의하지 않으므로 4개 변수만으로 추정한 결과 상위랭킹(성공, UP, 알파벳 순위로 DN 이후 이므로 성공) 확률은 99.96%이다.

```
predict(lpga.logit0,newdata=data.frame(그린_적중률=65,평균_퍼팅수=28,샌드_회수
=1.5,샌드_세이브=40),type ="response")
```

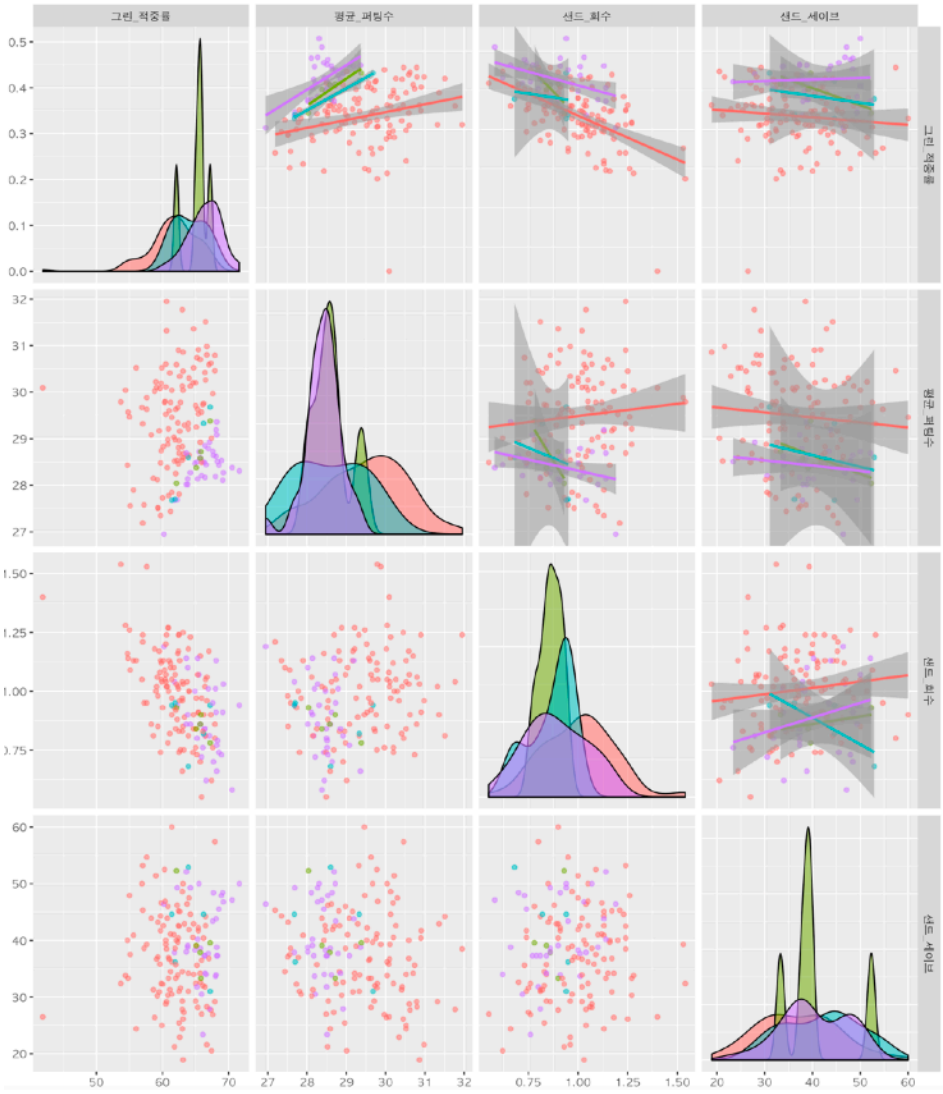
```
> predict(lpga.logit0,newdata=data.frame(그린
 ="response")
        1
0.9995064
```

(10) 판별결과 시각화 : 활용 및 해석 방법은 이전 강의노트 참고

```
lpga$classResult<-paste(lpga$ranking,'-',lpga$class0,sep='')
table(lpga$classResult)
library(GGally)
ggpairs(lpga[,4:7],aes(colour=lpga$classResult,alpha=0.4),upper=list(continuous='smooth'))
```

분류 결과를 범주화 하였음. DN-DN, UP-UP 정분류 결과임

```
> table(lpga$classResult)

DN-DN DN-UP UP-DN UP-UP
  112     5     5    35
```



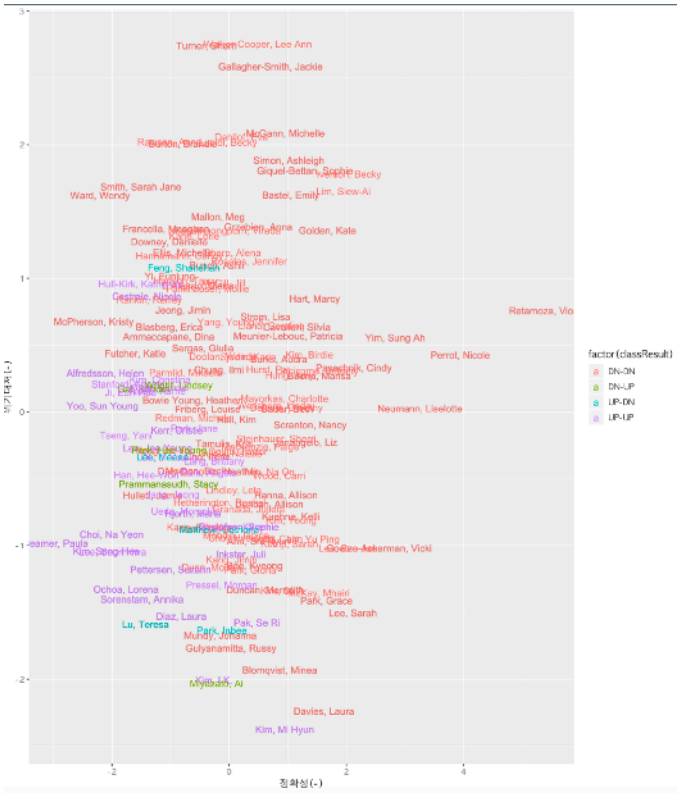
(11) 주성분 활용 판별결과 시각화

```
lpga.pca=prcomp(lpga[,4:7],scale.=T) #상관계수행렬 이용
lpga.pca$rotation #부하 출력
lpga.logit.pca<-cbind(lpga,lpga.pca$x)
library(ggplot2)
ggplot(lpga.logit.pca,aes(PC1,PC2,label=골퍼))+
  geom_text(check_overlap = TRUE)+
  geom_text(aes(colour = factor(classResult)))+
  xlab('정확성(-)')+ylab('퍼팅능력(-)')
```

제1주성분 : 그린적중율과 샌드회수는 반대 개념으로 정확성 지표로 볼수 있음 (- 정확성)

제2주성분 : 샌드세이브와 퍼팅 수는 적을수록 좋은 지표(- 위기대처 능력)

```
> lpga.pca$rotation #부하 출력
                PC1         PC2       PC3        PC4
그린_적중률 -0.680590378  0.13803045 0.3022330  0.6529927
평균_퍼팅수  0.180894150  0.69300418 0.6484046 -0.2580582
샌드_회수    0.709976133 -0.04144984 0.1217874  0.6923753
샌드_세이브  0.002816899 -0.70638142 0.6880367 -0.1662011
```



3) 사례연구2 : LPGA 데이터 (상금 랭킹 **상, 중, 하** 순서형)

종속변수 : 상위랭킹 여부(상금순위 40위 이내='상', 41-80위 = '중', 81위 이상 '하')

판별변수(독립변수) : 6개-(비거리, 페어웨이 안착율, 샌드(벙커) 세이버, 그린 적중률, 버팅 수)

```
lpga<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/
lpga2008.csv', fileEncoding = "utf-8")
lpga$group[rank(-lpga$상금)<=40]<-'TOP'
lpga$group[rank(-lpga$상금)>40 & rank(-lpga$상금)<=80]<-'MID'
lpga$group[rank(-lpga$상금)>80]<-'LOW'
```

```
> head(lpga)
               골퍼 평균_비거리 페어웨이_안착율 그린_적중률 평균_퍼팅수 샌
1     Ahn, Shi Hyun       249.4           64.6        61.2
2 Alfredsson, Helen       253.8           62.7        68.2
3 Ammaccapane, Dina       246.3           70.2        64.6
4       Bader, Beth       249.1           64.1        61.2
5       Bae, Kyeong       244.0           62.4        60.7
6     Baena, Marisa       254.2           64.7        60.9
  샌드_세이브  상금 참가_라운드수 group
1        34.5  6063            50   MID
2        38.8 19343            74   TOP
3        40.5  1873            50   LOW
4        41.1  1212            65   LOW
```

(1) 순서형 ordinal 로짓모형(추정)

그룹 $g = LOW, MID, TOP$

모형 : $logit(P(Y \le g)) = log\dfrac{P(Y \le g)}{P(Y > g)} = -X\underline{b}$

회귀계수가 -(음)인 것에 주의. 단 절편은 원래 추정 계수 그대로 (**R 코드만 그렇다**)

```
library(MASS)
lpga.lm<-polr(factor(group)~평균_비거리+페어웨이_안착율+그린_적중률+평균_퍼팅수
+샌드_회수+샌드_세이브,data=lpga, Hess=TRUE)
summary(lpga.lm)
```

로짓분석 모형 추정결과의 요약 리스트

- 유의확률이 출력되지 않음, 다음 방법에 의해 유의확률 출력이 가능하지만 t-값이 2미만이 경우 유의하지 않다고 하면 됨

- 페어웨이 안착율(가장 유의하지 않음), 평균비거리, 샌드 회수 유의하지 않음
- 이진형 로짓모형과 같이 stepAIC() 함수 사용 가능

```
Coefficients:
                   Value Std. Error  t value
평균_비거리      -0.008491    0.02860 -0.29693
페어웨이_안착율 -0.002917    0.05626 -0.05185
그린_적중률       1.204150    0.18206  6.61388
평균_퍼팅수      -3.566397    0.46407 -7.68512
샌드_회수         2.501478    1.79250  1.39552
샌드_세이브       0.077351    0.03216  2.40516

Intercepts:
        Value       Std. Error t value
LOW|MID   -25.0513     0.0076 -3282.4651
MID|TOP   -21.1727     0.6179   -34.2666
```

(2) 유의변수 선택

```
stepAIC(lpga.lm,direction='both')
```

- step() 함수 사용가능, both 대신 forward, backward 사용가능

최종 단계에 선택된 변수에 의한 추정모형이 출력된다.

```
Coefficients:
그린_적중률 평균_퍼팅수    샌드_회수 샌드_세이브
 1.19114047 -3.55534836  2.38915286  0.07779749
```

(3) 유의변수 모형추정

AIC 방법으로 샌드 회수도 유의하였으나, t-값이 여전히 2 이하였으므로 최종 3개 판별변수만 사용하였음

```
lpga.lm0<-polr(factor(group)~그린_적중률+평균_퍼팅수+샌드_세이
브,data=lpga,Hess=TRUE)
summary(lpga.lm0)
exp(coef(lpga.lm0)) #odds ratio
```

AIC (Akai keInformation Criterion) $AIC = 2(p - ln(Likihood))$(충분히 작아지는 변수개수) 기준 변수선택이므로 유의하지 않은 변수(샌드회수)가 출력된다.

```
Coefficients:
               Value Std. Error t value
그린_적중률  1.10890    0.15824   7.008
평균_퍼팅수 -3.37338    0.46934  -7.187
샌드_세이브  0.08942    0.03112   2.874

Intercepts:
        Value    Std. Error t value
LOW|MID -25.1032   8.3241    -3.0157
MID|TOP -21.2708   8.1518    -2.6093
```

*) 추정 회귀계수에 - 붙여야 함(단, 절편은 의미없음)

(a) 추정 로짓 모형 및 해석

$$log\frac{P(Y=LOW)}{P(Y\geq MID)} = -25.1 - 1.108*Gr + 3.37*Putt - 0.089*Sand$$

- 하위랭킹과 중상 랭킹의 오즈비 : (하위 랭킹 선수는 중상 랭킹 선수에 비해)그린 적중율 -낮고, 샌드 세이브 -낮고, 퍼팅수 +많을 가능성 높음

$$log\frac{P(Y\leq MID)}{P(Y=TOP)} = -21.27 + 1.108*Gr - 3.37*Putt + 0.089*Sand$$

- (중상 랭킹 선수는 랭킹 상 선수에 비해)그린 적중율 -낮고, 샌드 세이브 -낮고, 퍼팅수 +많을 가능성 높음

(b) 오즈비 해석

- 상금 랭킹이 높은 집단은 낮은 집단에 비해 그린 적중률이 3배 높다.
- 상금 랭킹이 높은 집단은 낮은 집단에 비해 샌드 세이브 성공율은 1.09배 높다.
- 상금 랭킹이 높은 집단은 낮은 집단에 비해 평균 퍼팅 수는 29(=.034)배 적다.

```
> exp(coef(lpga.lm0)) #odds ratio
그린_적중률 평균_퍼팅수 샌드_세이브
 3.03103640  0.03427371  1.09354258
```

(4) 사후확률 추정결과

추정결과는 $fitted.values 저장되어 있음

```
head(lpga.lm0$fitted.values)
```

```
> head(lpga.lm0$fitted.values)
          LOW         MID          TOP
1 0.029663118 0.55564577 0.4146911163
2 0.005721964 0.20420593 0.7900721102
3 0.819944269 0.17532208 0.0047336464
```

(a) 정분류 결과표

```
posterior<-data.frame(lpga.lm0$fitted.values) #making dtat frame
posterior$MAX<-apply(posterior,1,max,na.rm=TRUE)
posterior$class[posterior$MAX==posterior$TOP]<-'TOP'
posterior$class[posterior$MAX==posterior$MID]<-'MID'
posterior$class[posterior$MAX==posterior$LOW]<-'LOW'
lpga.class<-cbind(lpga,posterior[-4])
head(lpga.class)
```

사후확률 계산 결과와 가장 큰 사후확률 그룹으로 분류한 결과를 원 데이터 lpga 자장하였음

```
          LOW         MID           TOP class
1 0.029663118 0.55564577 0.4146911163   MID
2 0.005721964 0.20420593 0.7900721102   TOP
3 0.819944269 0.17532208 0.0047336464   LOW
```

```
lpga.class.ct<-table(lpga.class$group,lpga.class$class)
prop.table(lpga.class.ct,1) #행 퍼센트,분류표
sum(diag(prop.table(lpga.class.ct))) #정분류
```

(그린_적중률+평균_퍼팅수+샌드_세이브) 3개 판별변수를 이용하여 분류한 결과 정분류 비율이 81.5%이다. (샌드 회수까지 포함한 경우 정분류 비율은 80.9%로 낮아짐)

본 로짓판별규칙은 중위랭킹 정분류 비율은 62.5% 매우 낮다.

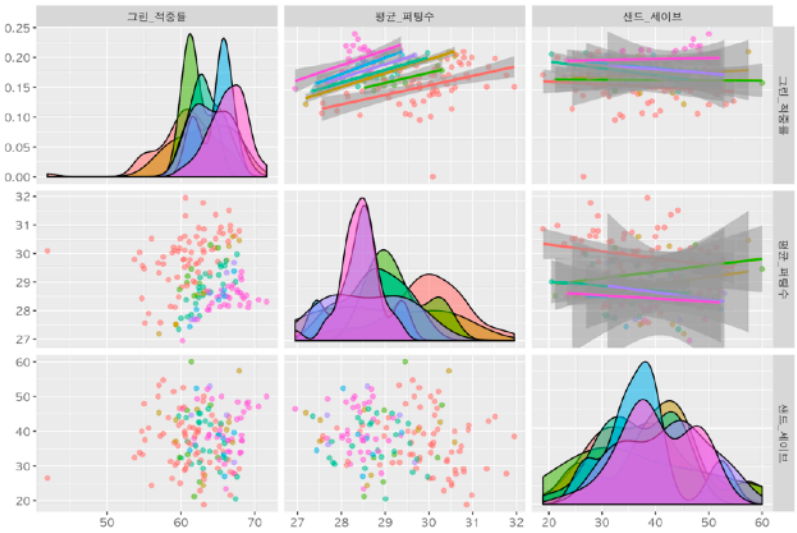
```
> prop.table(lpga.class.ct,1) #행 퍼센트,분류표

            LOW       MID       TOP
  LOW 0.8831169 0.1168831 0.0000000
  MID 0.2000000 0.6250000 0.1750000
  TOP 0.0000000 0.1250000 0.8750000
> sum(diag(prop.table(lpga.class.ct))) #정분류
[1] 0.8152866
```

(5) 판별결과 시각화 (활용 및 해석 방법은 이전 강의노트 참고)

```
lpga.class$classResult<-paste(lpga.class$group,'-',lpga.class$class,sep='')
table(lpga.class$classResult)
library(GGally)
theme_update(text=element_text(family="AppleGothic"))
ggpairs(lpga[,c(4,5,7)],aes(colour=lpga.class$classResult,alpha=0.4),upper=list
(continuous='smooth'))
```

분류 결과를 범주화(변수명 classResult) 하였음. TOP-TOP, TOP-MID, ...

```
> table(lpga.class$classResult)

LOW-LOW LOW-MID MID-LOW MID-MID MID-TOP TOP-MID TOP-TOP
     67      10       8      25       7       5      35
```



(6) 주성분 활용 판별결과 시각화

```
lpga.pca=prcomp(lpga.class[,c(4,5,7)],scale.=T) #상관계수행렬 이용
summary(lpga.pca) #변동(누적)설명기여율 출력
lpga.pca$rotation #부하 출력
lpga.logit.pca<-cbind(lpga.class,lpga.pca$x)
```

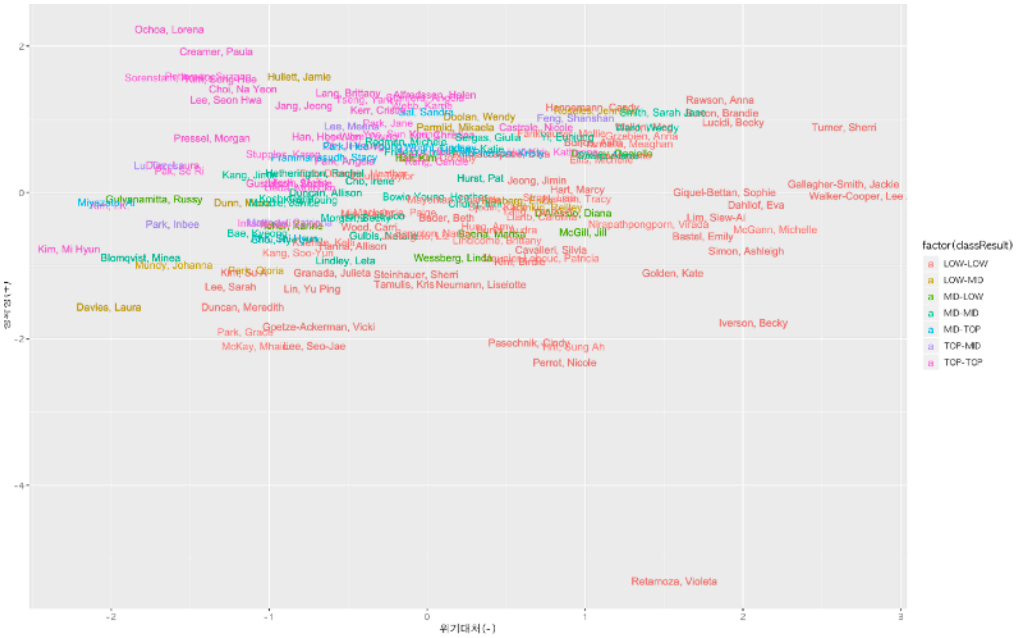
제1주성분 : 샌드 세이브를 잘하면 위기 대체 능력 높고 못하면 퍼팅 수가 맞아짐 (- 위기대처)

제2주성분 : 그린 적중률이 높다는 것은 티샷이 정확하다는 것을 의미함 (정확성)

```
> lpga.pca$rotation #부하 출력
                      PC1         PC2         PC3
그린_적중률   0.006875151 0.99101500   0.1335739
평균_퍼팅수   0.707423655 0.08958851 -0.7010889
샌드_세이브 -0.706756326 0.09931345 -0.7004515
```

```
library(ggplot2)
ggplot(lpga.logit.pca,aes(PC1,PC2,label=골퍼))+
  geom_text(check_overlap = TRUE)+
  geom_text(aes(colour = factor(classResult)))+
  xlab('위기대처(-)')+ylab('정확성(+)')
```

Park, INbee - 상위랭크 선수이나 로짓판별 결과 중위랭킹으로 오분류, 이유는 대처 능력은 상위랭킹 수준이나 정확성이 중간 수준이라 오분류 되었음.



(7) 새로운 선수 판별

이진형(UP 상금1~40위, DN) 로짓판별함수는 (샌드 회수=1.5 포함) 상위랭킹(UP) 확률은 99.96%였다. 순서형(TOP 상금1~40위, MID 상금41~80위, LOW상금81위~) 순서형로짓판별 결과 TOP 랭킹 선수로 분류

```
predict(lpga.lm0,newdata=data.frame(그린_적중률=65,평균_퍼팅수=28,샌드_세이브=40))
```

```
> predict(lpga.lm0,newdata=data.frame(그린_적중률=65,평균_퍼팅수=28,샌드_세이브=40))
[1] TOP
Levels: LOW MID TOP
```

Decision Tree 분류가지

1) 개념

- 의사결정나무 방법은 모집단 개체를 판별(Discrimination) 변수에 의해 계층적(단계적) 분류하는 방법으로 컴퓨터의 연산 능력에 의존하는 기법 : 빅데이터 판별분석 중 하나임
- 판별변수는 측정형(양적)은 물론 질적(범주형) 변수도 가능함
- 목표변수는 일반적으로 이진형(성공,실패 결과가 2개) : 컴퓨터 계산 능력의 발전으로 3개 이상의 범주형도 가능하나 해석 측면에서 이진형이 적절함
- 개체를 판별변수에 의해 분화(일반적으로는 2개 가지이나 3개 이상 가능) 가지치기를 계층적으로 시행하여 최종 노드의 가지(범주)는 교집합이 없음(동일 개체은 최종노드에 한 곳에만 속한다) -> 최초 개체수와 최종노드 개체 수는 동일하다.

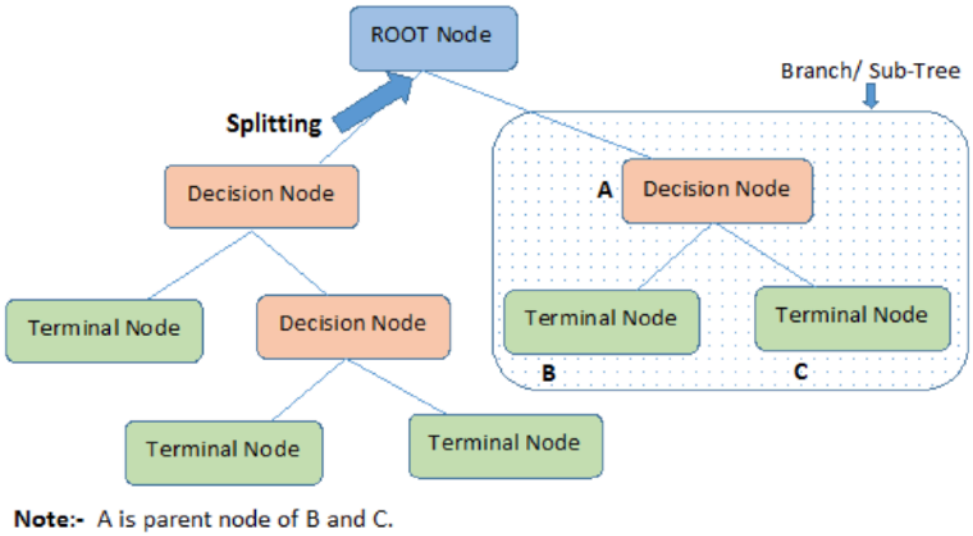
의사 결정 트리 관련 용어

루트 root 노드 : 전체 모집단 또는 샘플을 나타내며 두 개 이상의 동질 그룹(집단)으로 나뭇다.

분할 splitting : 2 개 이상의 서브(하위) 노드에 노드를 분할하는 처리이다.

의사 결정 decision 노드 : 하위 노드가 추가 하위 노드로 분할되면 이를 의사 결정 노드라 한다.

리프 / 터미널 노드 : 분할되지 않은 노드를 리프 또는 터미널 노드라 한다. 분할된 노드는 intermediate 노드라고 한다.



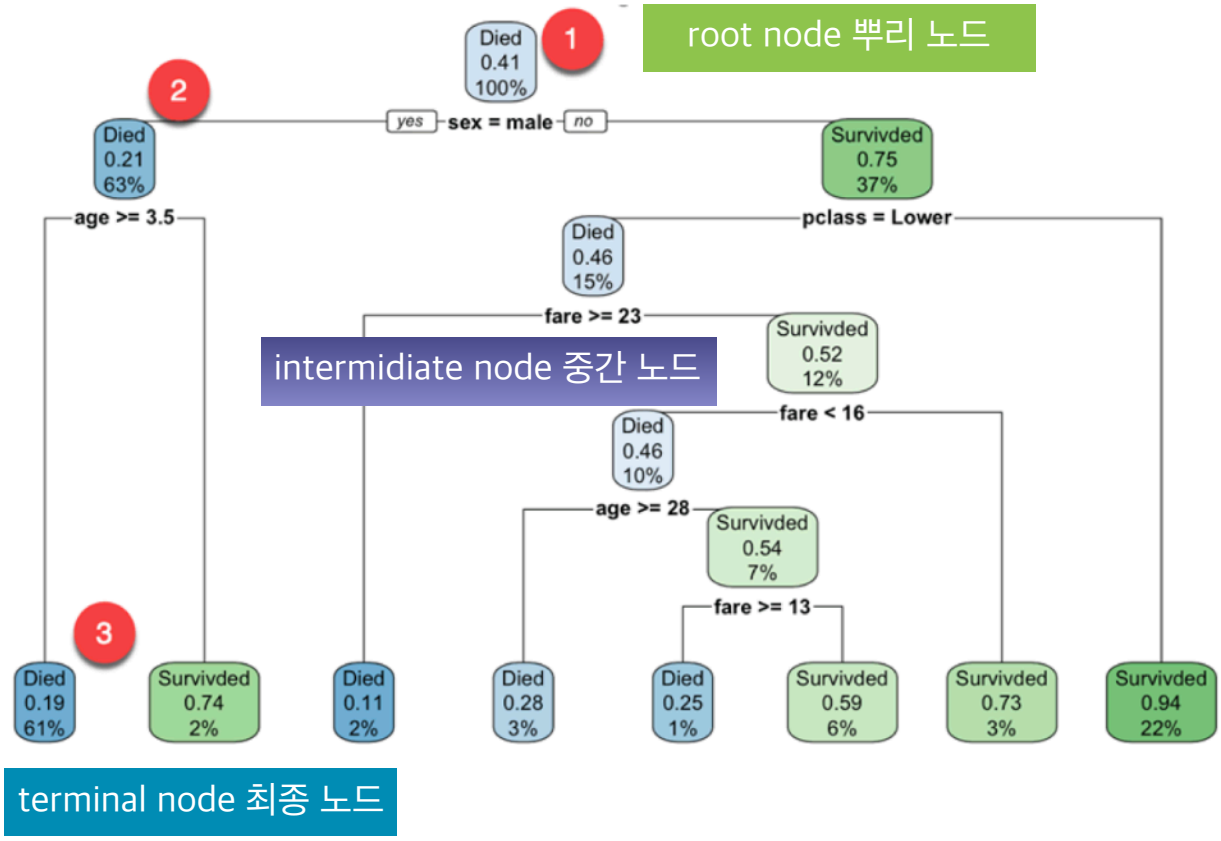
(출처) https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/2/

Prunning 가지치기: 의사 결정 노드의 하위 노드를 제거 할 때 프로세스이다. 반대의 분리 과정을 말할 수 있습니다.

가지 branch / 하위 트리 sub-tree : 전체 트리의 하위 섹션을 가지 또는 하위 트리라 한다.

부모 parent 및 자식 children 노드 : 상위계층 노드를 부모 노드, 부모 노드의 하위 노드를 자식 노드라 한다.

용어 관련 예제



(출처 https://www.guru99.com/r-decision-trees.html)

(1) 타이타닉 데이터의 승객 중 생존자 비율이 41%(0.41), 100%의 의미는 사망여부 분류 전이므로 전체 비율임. 본 분류에서 목표는 생존자 분류임

(2) 뿌리노드의 분류변수는 성병임. 남자는 전체 승객의 63%이며, 그 중 21%가 생존하였음. 뿌리 단계의 100%는 하위 노드에 상호배타적 분류되므로 63%(남자)+37%(여자) 판변변수 성별에 의해 상호배타 분리된다.

(3) 다음 중간노드(최종노드) 분류변수는 나이, 3.5세(*기준값 설정은 교차표 검정통계량의 값이 가장 큰 값 적용) 이상인 남자는 63% 중 2%(그러므로 남은 61%는 3.5세 미만)이고 19%의 생존율, 3.5세 미만은 74% 생존율을 보임

◆ 승객 중 여자 비율은 37% (오른쪽 첫 가지) - 낮은 운임 탑승객은 37% 중 15%(높은 운임 탑승객의 22%) 중 사망한 비율(왼쪽 가지가 YES)은 46%임, 여자이고 높은 운임 탑승자는 비율은 (0.37*0.22)%이고 그 중 생존 비율은 94%임.

의사 결정 트리의 종류

이진형 결정 트리

이진 목표 변수를 가진 결정 트리 (예) 타이타닉 승객 생존여부

범주형 결정 트리

목표 변수의 범주가 3개인 결정 트리 (예) 은행 대출기업 기한 만료 시 "완납", "기한연장", "부도"

연속 변수 결정 트리

의사 결정 트리에는 목표변수가 연속형인 경우 - 목표변수를 구간화(가능하면 이진형) 하여 범주형(이진형) 의사결정나무 사용하는 것이 적절함

장단점

- 의사결정나무는 분류와 예측판별에 강력하고 사용 빈도 많은 도구이다.
- 데이터가 가진 숨은 내재적 구조를 파악하여 이해하고 활용하기 쉬운 규칙을 도출한다.
- **이해하기 쉬운** : 의사 결정 트리 결과를 읽고 해석하기 위해 통계 지식이 필요하지 않습니다. 그래픽 표현은 매우 직관적이며 사용자는 자신의 가설을 쉽게 연관시킬 수 있다.
- **데이터 탐색에 유용 :** 의사 결정 트리는 가장 중요한 변수와 둘 이상의 변수 간의 관계를 식별하는 가장 빠른 방법 중 하나이다. 의사 결정 트리를 사용하여 대상 변수를 예측할 수있는 더 강력한 기능을 갖춘 새로운 변수 / 기능을 만들 수 있습니다. 데이터 탐색 단계에서도 사용할 수 있습니다.
- **데이터 유형은 제약 조건이 아니다** . 숫자 및 범주 변수를 모두 처리 할 수 있습니다.
- 비모수적 **방법 :** 의사 결정 트리는 비모수 적 방법으로 간주한다. 이는 의사 결정 트리에 공간 분포 및 분류 자 구조에 대한 가정이 없음을 의미한다.

❖ 단점

- 초과 적합 overfit **:** 초과 적합은 의사 결정 트리 모델에서 가장 실용적인 단점이다.
- 연속 변수에 적합하지 않음 : 연속 숫자 변수로 작업하는 동안 의사 결정 트리는 다른 범주의 변수를 분류 할 때 정보를 잃습니다.

2) 절차

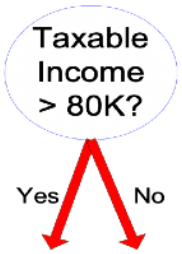
(1) 목표집단(뿌리노드 결정)과 판별변수 결정 which to start?

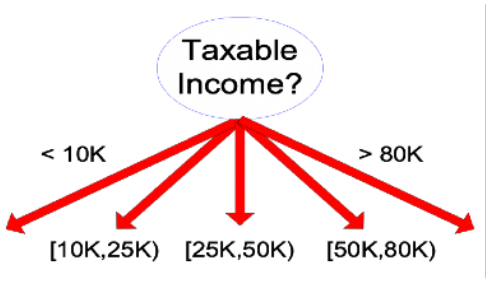
- 목표집단은 판별하려는 모집단의 그룹이며, 가능하면 2개 집단이 해석과 활용이 용이하다. -> **root node 결정 - 이진형, 범주형**
- 판별변수는 활용 가능한 모든 요인을 사용하는 것이 좋다. (빅데이터 소프트웨어 발전)

고객이 보험 회사에 갱신 보험료를 지불 할 것인지 예측하는 데 문제가 있다고 가정합니다 (예 / 아니요). 여기서 우리는 고객의 소득이 중요한 변수이지만 보험 회사가 모든 고객에 대한 소득 정보를 가지고 있지 않다는 것을 알고 있습니다. 이제 이것이 중요한 변수라는 것을 알았으므로 직업, 제품 및 기타 다양한 변수를 기반으로 고객 수입을 예측하는 의사 결정 트리를 작성할 수 있습니다. 이 경우 연속 변수의 값을 예측합니다.

(2) 계층적 노드(판별변수) 결정 및 노드 가지수 결정 which node to proceed

연속형 변수의 가지 수





(3) 노드 수 결정 when to stop

- 분리기준 split criterion, 정지 규칙 stopping rule
- 가지치기 : 오분류 가능성이 높아지는 경우 노드를 정지하게 됨
- 그러면 언제까지 나무모형을 성장시킬 것인가 ? 너무 큰 나무모형은 자료를 과대적합 하고 반대로 너무 작은 나무모형은 자료를 과소적합할 위험
- 일반적으로 사용되는 방법은 마디에 속하는 자료 가 일정 수(가령 5) 이하일 때 분할을 정지 (thumb rule)

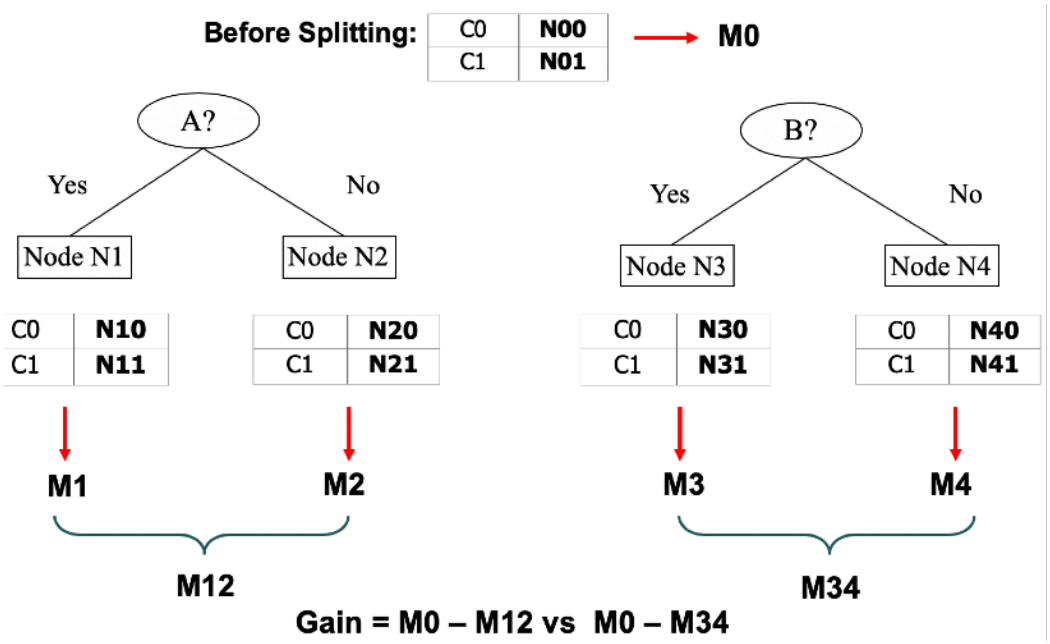
분산분석(회귀분석) 개념

- 판별변수가 범주형이면 교차분석 F-통계량 Classification tree
- 판별변수 측정형이면 범위의 값을 이분화 하여 교차분석을 반복하면서 F통계량을 가 장 크게 하는 설명변수 값을 찾는다. - Regression tree

3) 노드, 가지 결정

동질성, 불순도, 정보획득

- 노드의 동질성 homogeneous : 구분된 가지에 속한 개체들 중 동일 목표 그룹 속한 개체 비율을 의미함
- 노드의 불순도(impurity) : 동질성과 반대 개념임
- 노드의 가지 분리를 위하여 판별변수의 범주 기준값을 결정하기 위하여 노드의 동일성을 높여 가는 과정을 Information Gain 정보 획득 과정이라 한다.



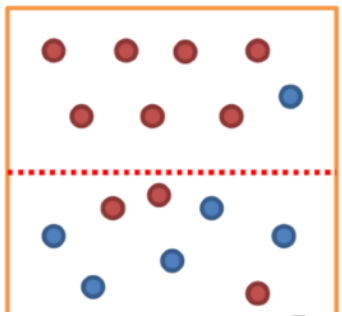
Entropy 엔트로피 (개체 동질성 척도) - 노드 결정 원칙

(a) 정의

$Entropy(t) = -\sum_{j=1}^{s} P(j|t)log(P(j|t))$, 노드 t, j = 가지(범주 영역)

$P(j|t)$ = 노드 t의 하위 가지(범주) j에 속한 개체들의 상대빈도

계산 예제 https://ratsgo.github.io/machine%20learning/2017/03/26/tree/



$$Entropy(A) = -\frac{10}{16}\log_2(\frac{10}{16}) - \frac{6}{16}\log_2(\frac{6}{16}) \approx 0.95$$

붉은 점선으로 가지 분리하면, 노드 엔트로피 0.95에서 0.75로 떨어짐

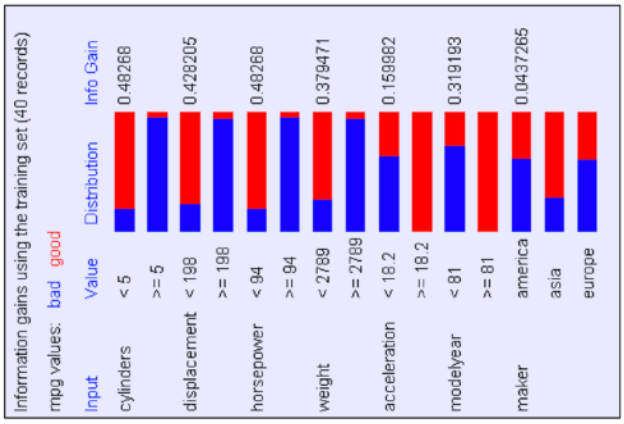
$$Entropy(A) = 0.5 \times \left(-\frac{7}{8}\log_2(\frac{7}{8}) - \frac{1}{8}\log_2(\frac{1}{8})\right) + 0.5 \times \left(-\frac{3}{8}\log_2(\frac{3}{8}) - \frac{5}{8}\log_2(\frac{5}{8})\right) \approx 0.75$$

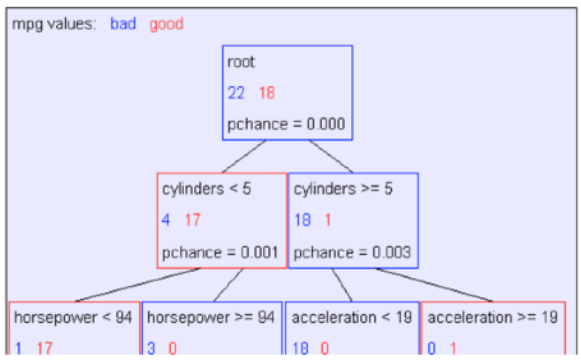
(b) 정보획득 (엔트로피)

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k}\frac{n_i}{n}Entropy(i)\right)$$ 부모 노드 p, 하위 자식 노드 i, 부모 노드 총 개체 수 n, 자식 i노드 개체수 n_i

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$ $$SplitINFO = -\sum_{i=1}^{k}\frac{n_i}{n}\log\frac{n_i}{n}$$

정보획득이 높은 변수와 분리기준(treshhold) 값을 결정한다. 실린더 먼저 결정





Gini Index (지니계수) - 불순도 척도, 노드 결정 원칙

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

노드 t 지니계수 :

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6

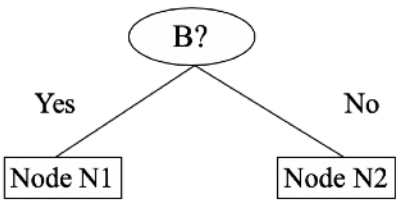
Gini = 1 – (1/6)² – (5/6)² = 0.278

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6

Gini = 1 – (2/6)² – (4/6)² = 0.444

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$



	Parent
C1	6
C2	6
Gini = 0.500	

Gini(N1)
= 1 – (5/6)² – (2/6)²
= 0.194

Gini(N2)
= 1 – (1/6)² – (4/6)²
= 0.528

	N1	N2
C1	5	1
C2	2	4
Gini=0.333		

Gini(Children)
= 7/12 * 0.194 +
5/12 * 0.528
= 0.333

Misclassification Rate (오분류율) - 노드 분류가지 오분류 정도, 노드 결정 원칙

$$Error(t) = 1 - \max_i P(i \mid t)$$

노드 t 오분류율 :

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Error = 1 – max (0, 1) = 1 – 1 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6

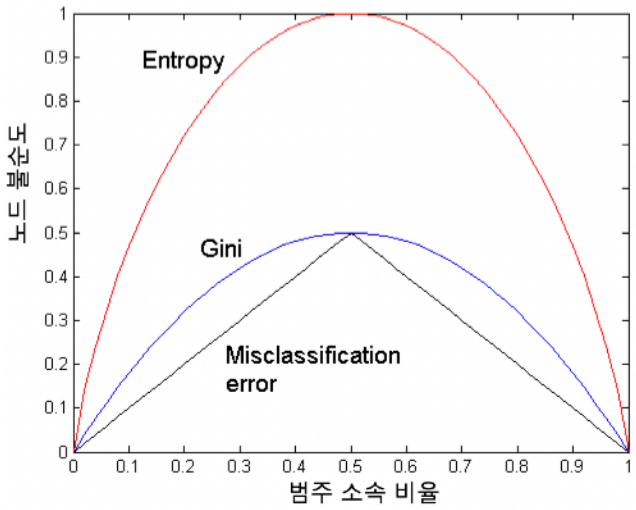
Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6

Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

자식노드 2개인 경우 범주 속할 비율과 불순도 관계



이슈사항

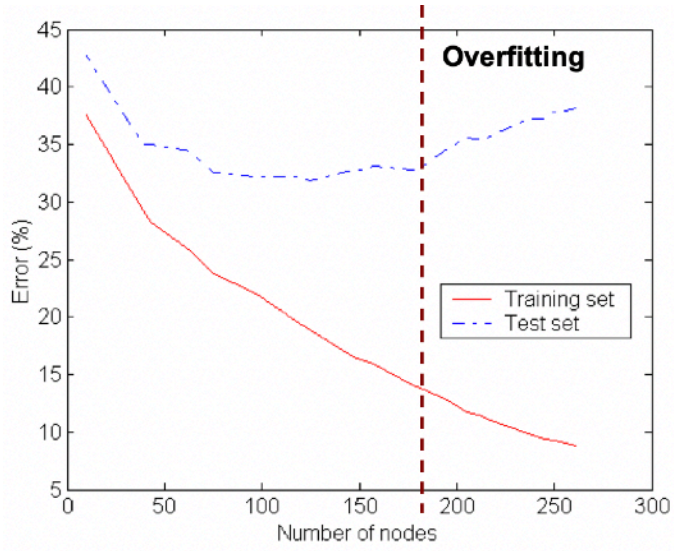
(a) 언제 노드 분기를 멈출 것인가?

표본 데이터 내 모든 개체가 하위 가지(범주)에 완전 분류되었을 때

(b) 과적합 overfitting - 해결 pruning

노드 수가 일정 수보다 증가하면 오분류 비율이 증가하는데 이를 과적합이라 한다. 과적합이 일어나기 전에 분기 가지치기 pruning을 적용하여 오분류 비율을 줄일 수 있음

post-pruning 방법과 pre-pruning 방법이 있음

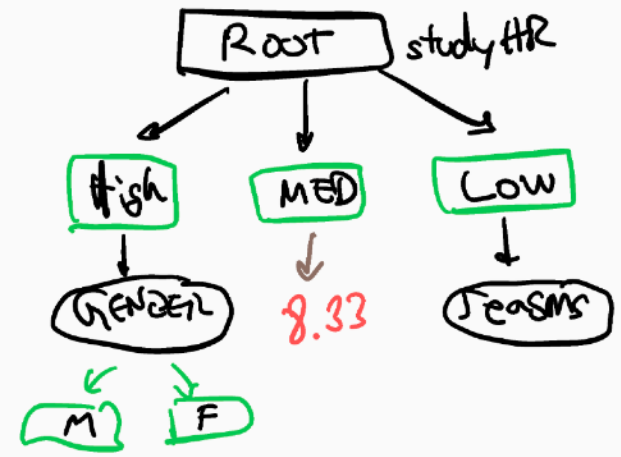


4) Decision Tree 의사결정나무 - 회귀분석

집단변수(목표변수)가 성공/실패(이진형), 순서형(상,중,하)이 아닌 측정형(예, 고등학생의 자기학습 공부시간 판별) 변수인 경우 회귀분석(종속변수=목표변수가 측정형) 개념의 의사결정나무 방법이 사용된다.

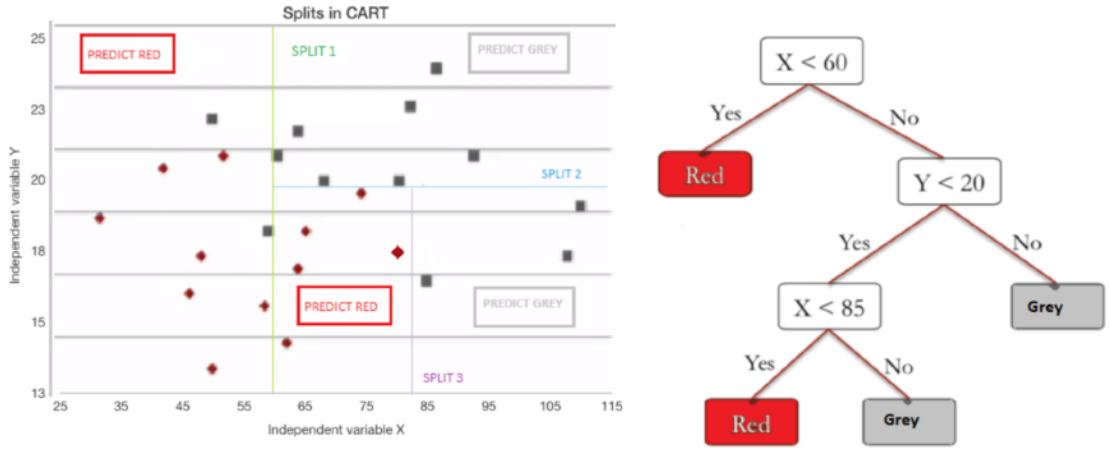
- 목표변수 Target : 고등학생들의 자기학습 공부시간'
- 판별(예측)변수 Predictor : 성별, 계절여부(4개 범주), 가구소득수준(3단계-상, 중, 하) - 목표변수가 측정형인 경우의 판별변수는 예측변수 이름을 주로 사용한다.

목표변수	판별변수		
공부시간	성별	가구소득	계절
12	남	상	봄
8	남	중	여름
6	남	하	가을
8	남	상	겨울
8	여	상	봄
8	여	중	여름
6	여	하	가을
9	여	중	겨울



(예시) root node는 소득수준에 의한 가지 분류되었고 소득수준 상은 성별 노드, 하는 계절노드로 2차 노드 가지분류되었으나 중은 더 이상의 하위 노드는 필요 없이 최종 평균 공부시간 5.5 시간으로 판별됨.

노드 가지, 하위 노드 구성도



https://medium.com/analytics-vidhya/a-guide-to-machine-learning-in-r-for-beginners-decision-trees-c24dfd490abb (출처)

SD 표준편차 개념

의사결정나무 는 뿌리노드부터 동질성(homogeneous, 유사성) 높은 개체를 배타적 가지로 분류해 개체를 판별하는 비모수적 방법이다. 표본 개체의 homogeneity는 표준편차로 계산된다. 만약 개체가 완전하게 동질하다면 표준편차는 0이다.

목표변수(Target='공부시간')의 표준편차 계산 $SD(T)$

표준편차$SD(y) = \sum (y_i - \bar{y})^2/(n)$=1.76 (평균 $\bar{y} = 8.125$) - 모집단 표준편차 계산이 적절

변동계수 계산 $CV = SD/\bar{y} = 1.76/8.125(*100)$ =21.7(%)

*) 변동계수가 10% 미만이면 더 이상 가지 치기(하위노드 구성)를 하지 않는다.

그러므로 만약 목표변수 변동계수가 10% 미만이면 의사결정나무 분석에 의한 판별은 적용 불가함

목표변수(Target)와 판별변수(Predictor) 범주별 표준편차 $SD(T,P) = \sum P(P \subset C)SD(P)$

	성별					가구소득			
	SD	빈도	확률(P)	SD*P		SD	빈도	확률	SD*P
남	2.18	4	0.5	1.090	상	1.89	3	0.38	0.707
여	1.09	4	0.5	0.545	중	0.47	3	0.38	0.177
			SUM	1.635	하	0.00	2	0.25	0.000
								SUM	0.884

SD 감소 reduction 계산

I) 목표변수의 최초 변동계수가 10% 미만, 이상을 판별하여 10% 이상인 경우 다음 단계를 실시한다.

II) 루트 노드 표준편처와 판별(예측변수) 표준편차의 차이를 계산한다. $SD(T) - SD(T,P)$

SD(T)-SD(T,P)	
성별	0.108
소득수준	0.600

III) 표준편차의 감소가 가장 큰 판별변인을 선택하여 하위 노드(leaf node 가지 노드)를 구성한다. -> 소득수준의 차이가 성별보다 크므로 감소가 많은 소득수준을 1차 노드로 구성한다.

IV)소득수준의 가지의 변동계수가 10% 이상이면 하위 노드를 위와 동일한 방법으로 구축하지만 그렇지 않은 가지치기를 멈춘다. 위의 예제에서 소득수준이 중인 경우 표분편차(위의 표에서 0.47), 평균은 8.33 이므로 변동계수는 5.7%이다. 변동계수가 10% 미만이므로 분기를 멈추고 소득수준 중의 평균 공부시간 8.33을 적으면 된다.

사례분석 : 타이타닉 생존자 판별 (의사결정나무 방법)

데이터 속성

- survival - Survival (0 = No; 1 = Yes)
- class - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- name - Name
- sex - Sex
- age - Age
- sibsp - Number of Siblings/Spouses Aboard
- parch - Number of Parents/Children Aboard
- ticket - Ticket Number
- fare - Passenger Fare
- cabin - Cabin
- embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
- boat - Lifeboat (if survived)
- body - Body number (if did not survive and body was recovered)

```
titanic<-read.csv('http://wolfpack.hnu.ac.kr/Stat_Notes/adv_stat/DATA/
titanic_original.csv')
names(titanic); head(titanic,3); dim(titanic); str(titanic)
```

```
> names(titanic)
 [1] "pclass"    "survived"  "name"      "sex"       "age"       "sibsp"     "parch"
 [8] "ticket"    "fare"      "cabin"     "embarked"  "boat"      "body"      "home.dest"
```

행(승객) 1310, 열은 14개 변수, str()함수는 데이터프레임 구조(structure) 보기, int=정수, chr=문자, num=실수, factor=범주형 요인(분산분석의 요인)

```
> dim(titanic)
[1] 1310   14
> str(titanic)
'data.frame':   1310 obs. of  14 variables:
 $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
 $ name     : Factor w/ 1308 levels "","Abbing, Mr. Anthony"
 ...
 $ sex      : Factor w/ 3 levels "","female","male": 2 3 2 3
 $ age      : num  29 0.917 2 30 25 ...
 $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
 $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
```

```
library(DataCombine)
titanic.clean<-DropNA(titanic,Var=c('survived'))
dim(titanic.clean)
table(titanic.clean$survived) #1=survived 0=dead
```

DropNA() 함수는 설정 변수에 NA(결측값)이 있으면 제외하는 함수로 DataCombine 패키지에 있음 - 1개 결측치가 있어 삭제 되어 최종 1,309 탑승객 정보 있음, 그중 500명 생존하여 생존율 38.2%(500/1309)

```
> table(titanic.clean$survived) #1=survived 0=dead

  0   1
809 500
```

판별변수 개별분석

범주형 판별변수

성별

```
titanic.ct.sex<-table(titanic.clean$survived,titanic.clean$sex)
titanic.ct.sex
prop.table(titanic.ct.sex,2) #열 퍼센트 행=1
chisq.test(titanic.ct.sex[,-1])
```

```
> titanic.ct.sex

        female male
  0   0    127  682
  1   0    339  161
> prop.table(titanic.ct.sex,2) #열 퍼센트 행=1

        female      male
  0  0.2725322 0.8090154
  1  0.7274678 0.1909846
> chisq.test(titanic.ct.sex[,-1])
```

- sex 변수에 공백이 있어 교차분석, 카이제곱 검정에서는 교차표의 1열을 삭제하였음
- 남자는 19.1% 생존, 여자는 72.7% 생존하였고 카이제곱 검증 결과 유의하였으므로 남녀별 생존율 차이는 매우 유의함 highly significant

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  titanic.ct.sex[, -1]
X-squared = 363.62, df = 1, p-value < 2.2e-16
```

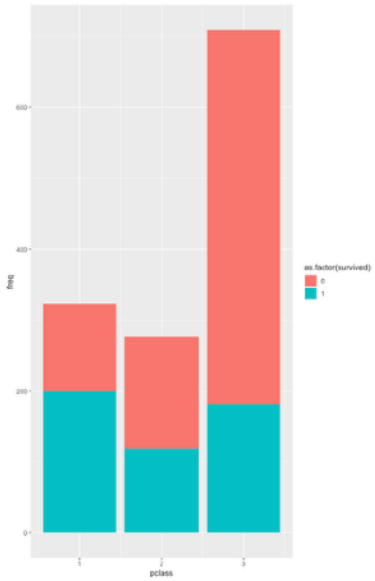
pclass 요금 : titanic.ct.pclass<-table(titanic.clean$survived,titanic.clean$pclass)

1등석 요금 생존자, 61.9%, 2등석 생존자 43%, 3등석 생존자 25.5%, 유의확률 <0.001 - 매우 유의한 차이

```
> prop.table(titanic.ct.pclass,2)

             1         2         3
  0 0.3808050 0.5703971 0.7447109
  1 0.6191950 0.4296029 0.2552891
```

```
> chisq.test(titanic.ct.pclass)

        Pearson's Chi-squared test

data:  titanic.ct.pclass
X-squared = 127.86, df = 2, p-value < 2.2e-16
```

```
library(ggplot2)
ggplot(data=titanic.count.pclass, aes(x=pclass, y=freq, fill=as.factor(survived)))
+ geom_bar(stat="identity")
```



측정형 판별변수

```
library(doBy)
summaryBy(age+fare+sibsp~survived,titanic.clean,na.rm=T)
t.test(age~survived,titanic.clean,na.rm=T)
t.test(fare~survived,titanic.clean,na.rm=T)
t.test(sibsp~survived,titanic.clean,na.rm=T)
```

생존자의 평균 나이 28세 어릴수록, 평균 요금 49 요금 많이 낸 승객이, 동반가족 수가 0.46명 적은 사람의 생존 가능성이 높음

```
> summaryBy(age+fare+sibsp~survived,titanic.clean,na.rm=T)
  survived age.mean fare.mean sibsp.mean
1        0 30.54537  23.35383  0.5216316
2        1 28.91823  49.36118  0.4620000
```

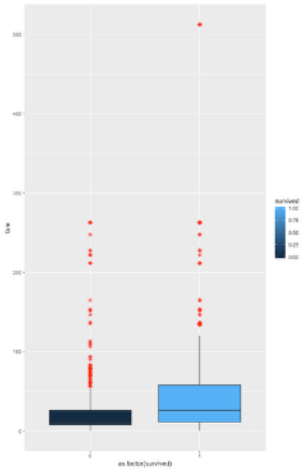
```
t.test(age~survived,titanic.clean,na.rm=T)
t.test(fare~survived,titanic.clean,na.rm=T)
t.test(sibsp~survived,titanic.clean,na.rm=T)
```

```
data:  age by survived
t = 1.7707, df = 868.26, p-value = 0.07696
```
나이는 약간 유의(유의수준 10%)

```
data:  fare by survived
t = -7.8889, df = 654, p-value = 1.284e-14
```
요금은 매우 유의

```
data:  sibsp by survived
t = 1.1371, df = 1298, p-value = 0.2557
```
동반가족수 유의하지 않음

```
library(ggplot2)
ggplot(data=titanic.clean, aes(x=as.factor(survived),
y=fare, fill=survived))
+geom_boxplot(outlier.colour="red",outlier.shape=8,
outlier.size=2)
```



의사결정 분석

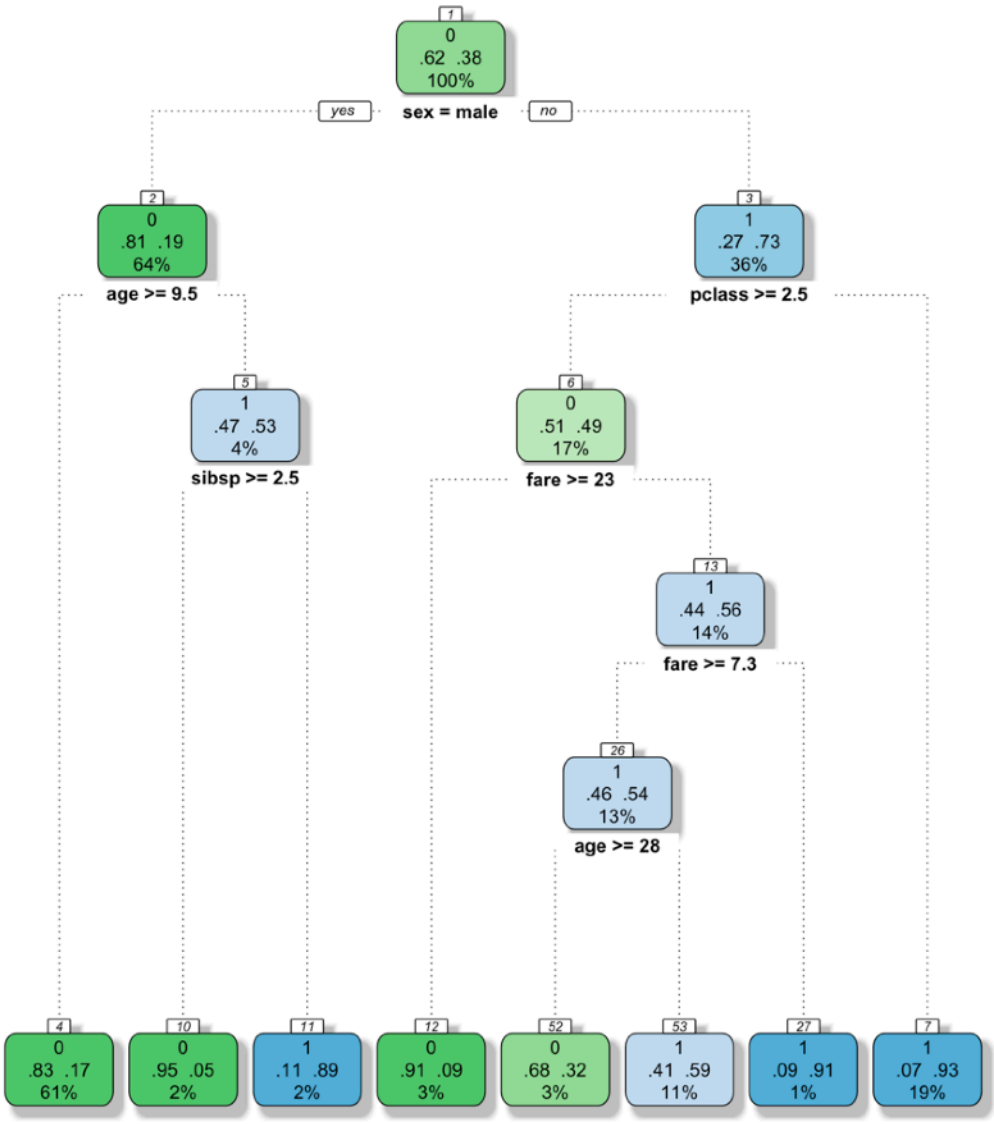
목표변수 : 승객의 생존여부 - 판별변수 : 요금등급, 성별, 나이, 동반 가족수, 요금, 선실이름

```
library(rpart)
titanic.tree<-rpart(survived~pclass+sex+age+sibsp+fare+embarked,
data=titanic.clean, method='class')
library(rattle); library(rpart.plot); library(RColorBrewer)
fancyRpartPlot(titanic.tree)
```

| 결과해석 | 모집단 (사망=809, 생존 509)

- Root 노드 : 생존비율 62%, 사망비율 38%
- 성별 판별 변인의 판별 능력이 가장 높아 최상위 가지 노드 구축 : 성별=남자 (예)는 왼쪽 가지 (2) 노드(승객비율 64%(남자 승객수 843명), (아니오) 오른쪽 가지 (3)노드 승객비율36%(여자승객수 466명)

```
> table(titanic.clean$sex)

female   male
     0    466    843
```



- 2번 가지 노드의 하위 가지노드의 판별변인은 나이이고 기준 값은 9.5세임. 4번 가지노드(남자승객이고 나이 9.5세이상) 미만자 비율은 61%(총 승객 중 남자&나이 9.5세 이상 승객비율)이고 사망 비율은 83%, 생존 비율은 17%임
- 5번 가지노드(남자이고 나이 9.5세 미만)인 승객 비율은 4%(총합이 64%가 아닌 65%인 이유는 반올림 결과임)은 다시 동반가족 수 2.5 이상/미만을 기준으로 가지노드 분류됨
- 10번 노드(남자, 나이9.5세미만,동반가족수2.5명이상) 비율은 2% - 생존비율 0.05%
- 11번 노드(남자, 나이9.5세미만,동반가족수2.5명미만) 비율은 2% - 생존비율 0.89%

의사결정 분석 정분류 표

```
titanic.predict<-data.frame(predict(titanic.tree,titanic.clean,type="class")) #판별결과는 0, 1로 구성된 열벡터임, 이를 데이터프레임으로 만듬(열변수 1개)
colnames(titanic.predict)[1] <-'Predict_survived' #데이터프레임 열변수 이름 설정
titanic.tree.predict<-cbind(titanic.clean,titanic.predict) #원데이터, 판별결과 합치기
titanic.tree.predict.ct<-
table(titanic.tree.predict$survived,titanic.tree.predict$Predict_survived)
prop.table(titanic.tree.predict.ct,1) #행 퍼센트 출력
sum(diag(prop.table(prop.table(titanic.tree.predict.ct,1)))) #정분류
```

모집단생존(1)-판별생존(1) 정분류 비율 69.8% 모집단사망(0)-판별사망(0) 정분류 비율 90.5%임 - 의사결정나무 방법는 사망자 정분류 비율이 매우 높음 (전체 정분류 비율은 80.1%)

```
> prop.table(titanic.tree.predict.ct,1) #열 퍼센트

             0          1
  0 0.90482077 0.09517923
  1 0.30200000 0.69800000
> sum(diag(prop.table(prop.table(titanic.tree.predict.ct,1)))) #정분류
[1] 0.8014104
```

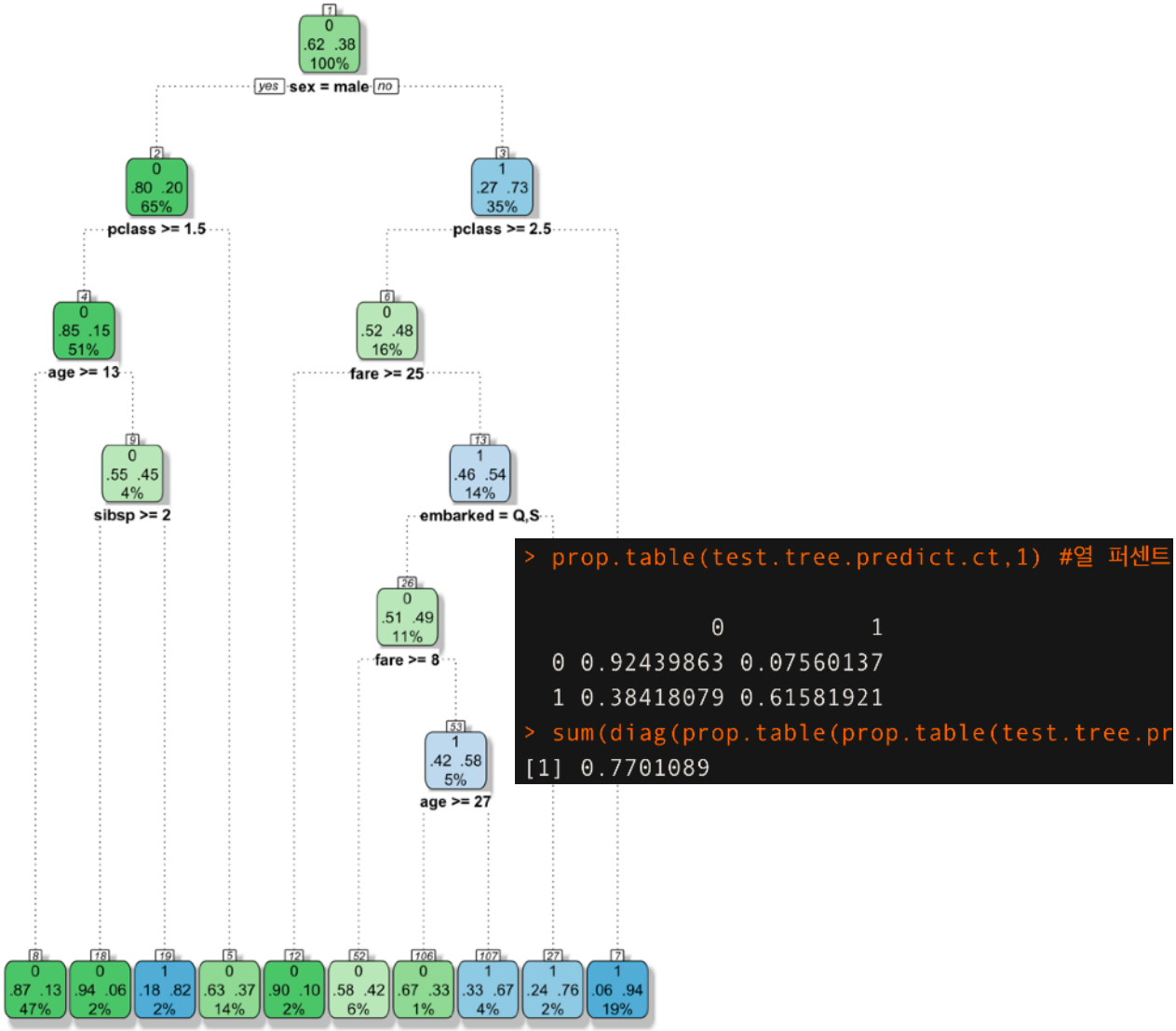
Train 데이터, Test 데이터 나눈 후 train 데이터

```
library(caTools)
set.seed(123)   #  set seed - train, test 동일 데이터 얻기
sample=sample.split(titanic.clean,SplitRatio = 0.7)
train=subset(titanic.clean,sample==TRUE)
test=subset(titanic.clean,sample==FALSE)
dim(titanic.clean); dim(train); dim(test)
```

```
> dim(titanic.clean); dim(train); dim(test)
[1] 1309   14
[1] 841  14
[1] 468  14
```

원 데이터를 Train, Test 데이터로 7:3 분리

```
train.tree<-
rpart(survived~pclass+sex+age+sibsp+fare+embarked,data=train,method =
“class")
fancyRpartPlot(train.tree)
test.predict<-data.frame(predict(train.tree,test,type="class"))
colnames(test.predict)[1] <-'Predict_survived'
test.tree.predict<-cbind(test,test.predict)
test.tree.predict.ct<-
table(test.tree.predict$survived,test.tree.predict$Predict_survived)
prop.table(test.tree.predict.ct,1) #행 퍼센트
sum(diag(prop.table(prop.table(test.tree.predict.ct,1)))) #정분류
```




```
> prop.table(test.tree.predict.ct,1) #열 퍼센트

             0          1
  0 0.92439863 0.07560137
  1 0.38418079 0.61581921
> sum(diag(prop.table(prop.table(test.tree.pr
[1] 0.7701089
```

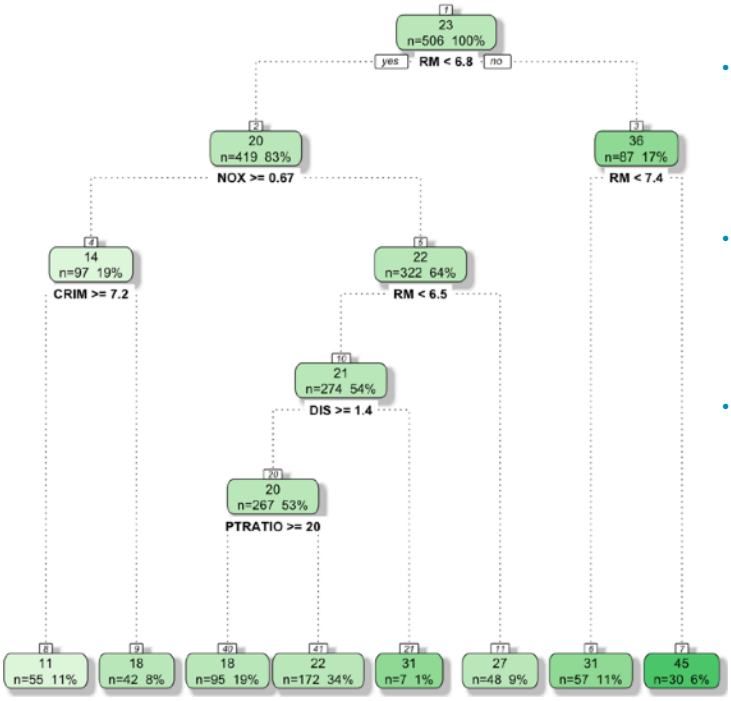
원 데이터의 70%만으로 의사결정 기법을 적용한 결과 루트 노드 판별 변인은 동일하나 남자 하위 노드는 나이가 아닌 요금 등급임 -> 모형 추정 데이터가 달라졌으니 당연함

Decision Tree - Regression 실증데이터 분석

데이터 & 분석 결과

LON and LAT are the longitude and latitude of the center of the census tract.
MEDV : 해당지역 개인 소유 주택 가격의 중위값(50%), measured in thousands of dollars. (목표변수)
CRIM is the per capita crime rate.
ZN is related to how much of the land is zoned for large residential properties.
INDUS is the proportion of the area used for industry.
CHAS is 1 if a census tract is next to the Charles River else 0
NOX is the concentration of nitrous oxides in the air, a measure of air pollution.
RM is the average number of rooms per dwelling.
AGE is the proportion of owner-occupied units built before 1940.
DIS is a measure of how far the tract is from centres of employment in Boston.
RAD is a measure of closeness to important highways.
TAX is the property tax per $10,000 of value.
PTRATIO is the pupil to teacher ratio by town.

```
boston<-read.csv('http://203.247.53.31/Stat_Notes/adv_stat/DATA/boston.csv')
str(boston)
library(rpart)
tree.boston=rpart(MEDV~LAT+LON+CRIM+ZN+INDUS+CHAS+NOX+RM+AGE
+DIS+RAD+TAX+PTRATIO,data=boston)
library(rattle); library(rpart.plot); library(RColorBrewer)
fancyRpartPlot(tree.boston)
```



- 목표변수가 측정형이므로 Terminal 최끝단 노드(가지) 위의 숫자는 목표변수 MEDV의 평균이다.
- 즉 RM(방의개수)>=6.8 이상이고, RM<7.4 미만인 지역의 주택의 중위값(MEDV) 평균은 31이다. (노드 31)
- 즉 RM(방의개수)>=6.8 이상이고, RM>=7.4 이상인 지역의 주택의 중위값(MEDV) 평균은 45이다.

분류 정확도

목표변수가 이진형, 범주형인 경우에는 판별 전후 집단 정분류, 오분류 결과를 활용하지만 MEDV 측정형 목표변수는 (실측값과 판별값)^2 (Sum of Squares of Errors 예측오차 제곱합)을 이용하면 된다.

```
pr<-predict(tree.boston, newdata=boston) #predcied value
head(pr,3)
sum((pr-boston$MEDV)^2) #SSE
```

```
> head(pr,3)
       1        2        3
27.27083 21.66919 31.34211
> sum((pr-boston$MEDV)^2)
[1] 10249.41
```

termimal 가지 내 추정값은 모두 동일하다. 즉 방의 개수 6.80이상, RM 7.40이상인 지역 개인 소유 주택 (중위값)가격은 31.3421로 모두 동일하다.

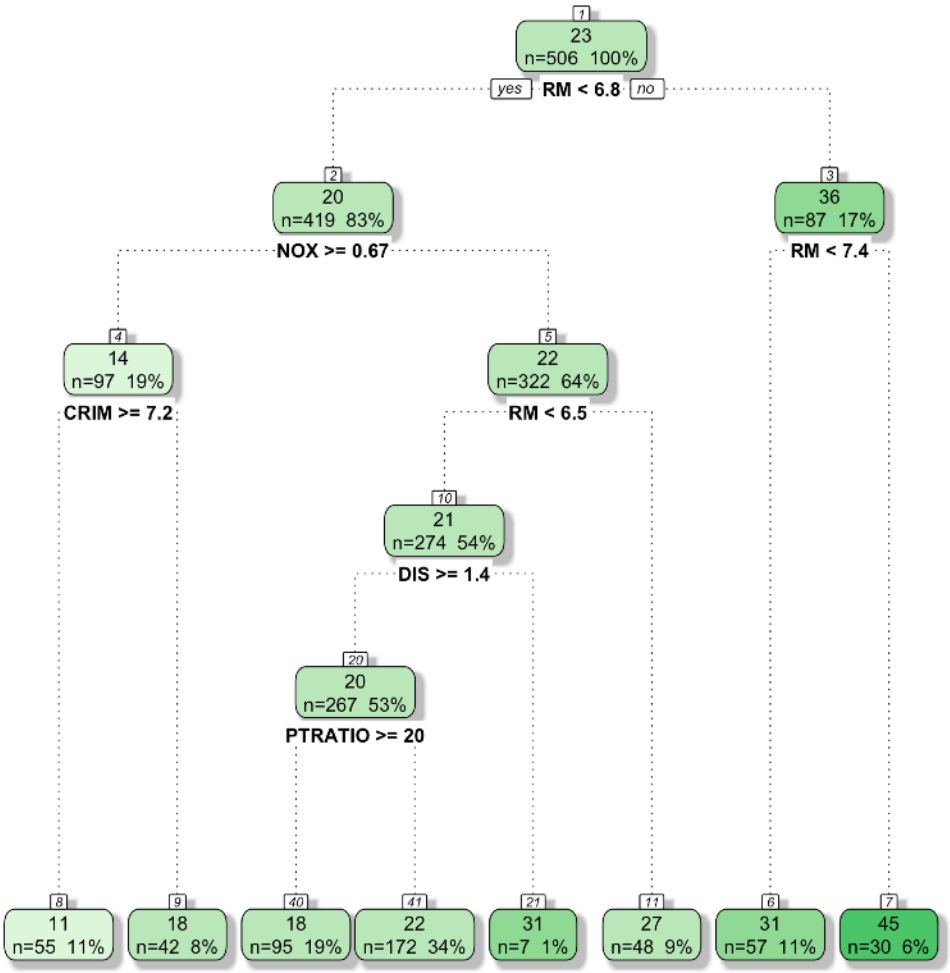
선형모형 적용 유의한 판별변수만 활용

```
lm.boston=lm(MEDV~LAT+LON+CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO,data=boston)
summary(lm.boston)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.680e+02  3.422e+02  -0.491 0.623743
LAT          2.250e+00  4.142e+00   0.543 0.587240
LON         -1.393e+00  3.809e+00  -0.366 0.714674
CRIM        -1.825e-01  3.608e-02  -5.059 5.97e-07 ***
ZN           4.001e-02  1.553e-02   2.577 0.010260 *
INDUS       -4.268e-02  6.968e-02  -0.613 0.540433
CHAS         3.468e+00  9.750e-01   3.557 0.000412 ***
NOX         -2.144e+01  4.498e+00  -4.766 2.47e-06 ***
RM           6.085e+00  3.987e-01  15.260  < 2e-16 ***
AGE         -4.559e-02  1.422e-02  -3.207 0.001426 **
DIS         -1.570e+00  2.358e-01  -6.659 7.39e-11 ***
RAD          2.548e-01  7.512e-02   3.392 0.000750 ***
TAX         -1.265e-02  4.216e-03  -3.001 0.002829 **
PTRATIO     -9.523e-01  1.540e-01  -6.185 1.30e-09 ***
```

매우 유의한 판별변수(***) 만 활용한 결과 동일한 의사결정 가지를 얻었고 판별능력 SSE도 동일하였다. 유의한 판별변수 선정한 후 의사결정 나무 활용하는 것이 적절함.

```
tree.boston0=rpart(MEDV~CRIM+CHAS+NOX+RM+DIS+RAD+PTRATIO,data=boston)
fancyRpartPlot(tree.boston0)

pr0<-predict(tree.boston0, newdata=boston); head(pr,3)
sum((pr0-boston$MEDV)^2)
```

```
> head(pr,3)
       1        2        3
27.27083 21.66919 31.34211
> sum((pr0-boston$MEDV)^2)
[1] 10249.41
```

노드 가진 분류 범주가 동일하므로 당연 가지 내 MEDV 평균 값은 동일함