Chapter 10. 판별분석

1. 판별분석의 개요

1. 판별분석이란

- 판별분석의 기본 개념
  - 피셔(Fisher)가 개발
  - 집단을 구분할 수 있는 설명변수를 통하여 집단 구분 함수식(판별식)을 도출하고, 소속된 집단을 예측하는 목적으로 사용
  - 등간척도나 비율 척도(메트릭)로 측정된 독립변수를 이용해 명목척도 또는 서열척도(메트릭)로 측정된 종속변수를 분류하는데 사용
- 판별식의 도출
  - 판별식 $z = a_{w_1}x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n$

- 판별분석의 종류
  - 종속변수의 집단수가 2개인 경우: 판별분석 (discriminant analysis)
  - 종속변수의 집단수가 3개 이상인 경우: 다중판별분석(multiple discriminant analysis: MDA)
- 판별분석의 가설
  - 귀무가설: 두 개 또는 그 이상의 집단의 평균이 동일하다.
  - 대립가설: 두 개 또는 그 이상의 집단의 평균이 동일하지 않다.

1. 판별분석이란?

⇒ 선형판별함수를 도출하여 자료를 2개 이상의 그룹으로 분류 시 분류오차를 최소로 하는 분류 기법이다.

⇒ 정량적 자료로 측정된 독립 변수들을 이용하여 명목 자료로 된 종속변수의 집단 구분을 예측하는데 이용 됨.

(예 1) 은행에서 대출 기업의

① 부채상태
② 자산상태
③ 영업실적

⇒ 정량적 자료 → 독립변수

를 이용하여

① 대출 가능 기업
② 대출 불가능 기업

⇒ 명목형 자료 → 종속변수

으로 나눔

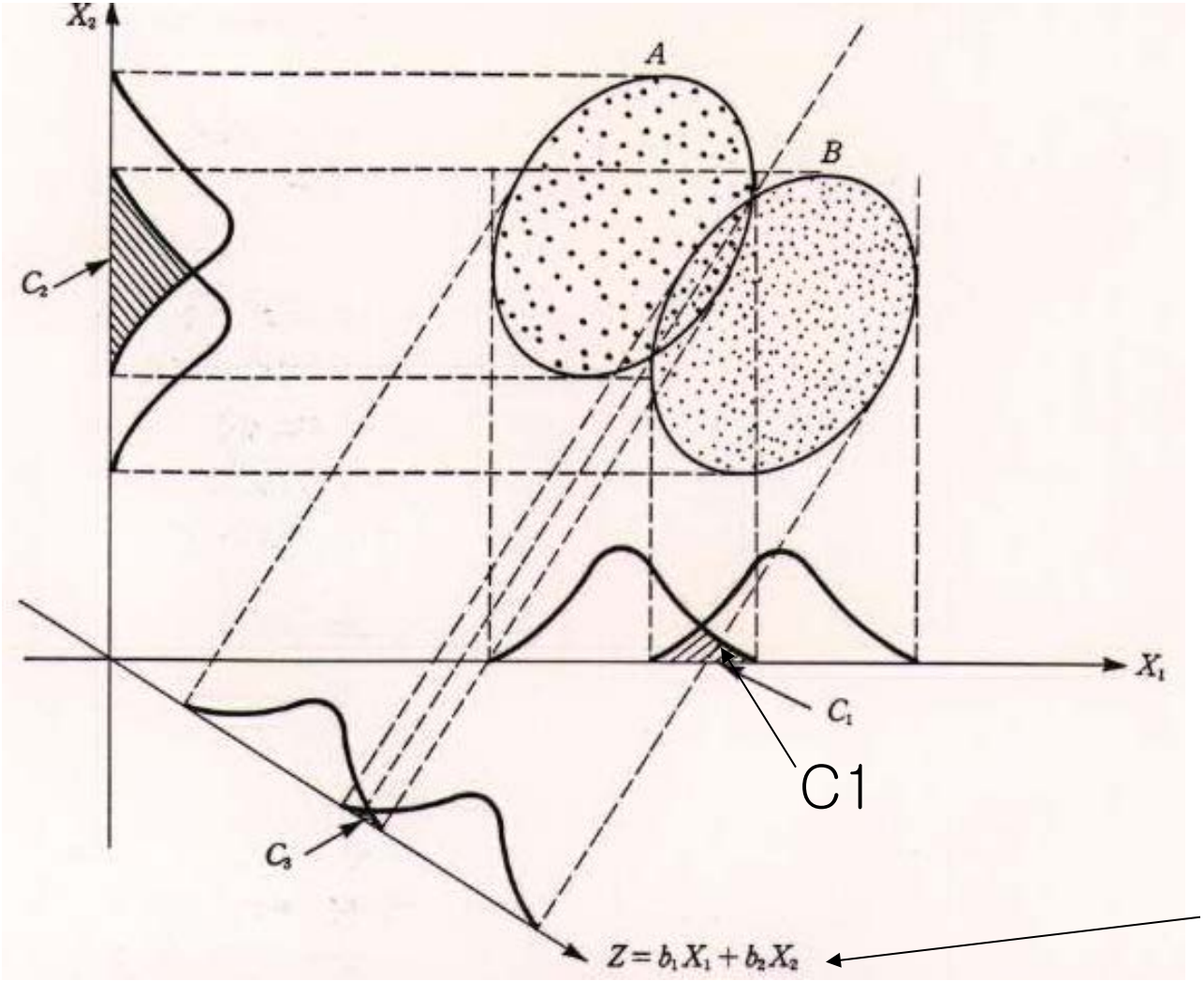
(예 2) 정량적 자료로 측정된 투표자의 특성에 따라
⇒ 어느 정당의 후보자를 지지할 것인가를 판단

<note> 판별분석의 목적
⇒ 이를 위해, 종속변수의 구분에 도움이 되는
① 독립변수의 선정
② 선정된 독립변수를 이용하여 판별함수를 도출
③ 판별 능력에 있어 독립 변수들의 상대적 중요도 평가
④ 판별함수의 판별 능력 평가
⑤ 새로운 판별 대상에 대한 예측력 평가

<note> 분류(classification)와의 차이점
1) 분류 : 대상이 몇 개의 그룹으로 나뉘어 지는가는 자료를 보기 전까지는 모름. – 예) 군집분석
2) 판별 : 존재하는 그룹의 수를 알고 있고, 새로운 대상이 어느 그룹에 속하는지를 결정함.

■ 기본 원리

⇒ 종속변수 그룹들 사이의 분산을 최대화 할 수 있는 새로운 축을 찾는 과정.



A와 B 두개의 그룹을

① X1 만으로 분류할 경우 → 오분류의 크기는 C1.

② X2 만으로 분류할 경우 → 오분류의 크기는 C2.

③ 오분류의 크기를 최소화 하기위해 → 새로운 함수 Z를 생성 함.

$Z=b_1X_1 + b_2X_2$

<note> 따라서,

① Z를 선형판별함수(linear discriminant function)이라 한다.

② 판별함수의 목적이

⇒ 종속변수의 그룹을 정확하게 분류하는데 있다면

→ 판별함수로부터 유의적인 판별력이 있는 독립 변수들을 선택한 다음

→ 분류를 위한 기준으로 판별함수로부터 계산한 판별득점(discriminant score)을 이용하는 방법을 이용한다.

<note> 도출할 수 있는 판별함수의 수

판별함수의 수 = Min { (그룹의 수 – 1) , 독립변수의 수 }

(예) 독립변수의 수가 5개이고, 3개의 그룹이면

Min {(3-1) , 5 } = 2(개)

2. 판별분석의 가정들

1) 독립 변수들의 결합확률분포는 다변량 정규분포이다.
2) 모집단에서 종속변수의 각 그룹 별로 독립변수들의 공분산 구조가 같다. → Box-M 검정을 이용

즉, $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_p = \Sigma$ 이다.

(예) X_1 = 부채상태, X_2 = 자산상태, X_3 = 영업실적 인 경우

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} \end{array} \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array}$$

대출가능기업 대출불가능기업

3) 각 그룹에 소속될 사전확률은 같다.
⇒ 각 그룹에 속한 표본의 수가 일단 같다고 가정하나
실제 연구에서는 지켜지기가 쉽지 않다.
→ 그러나, 위와 같이 가정하자.

<note> 참고
⇒ 표본의 크기가 매우 클 경우에는 가정이 별로 중요하지 않을 수도 있다.

3. 판별분석의 적용절차

● 제 1 단계 ⇒ 판별함수의 도출

1) 종속변수의 분류와 독립변수의 설정

① 종속변수의 선택 ⇒ 조사자는 전체 대상들을 몇 개의 그룹으로 나누어 분석할 것인가를 결정해야 한다.

<note> 이 때,

㉠ 각 그룹들이 상호 배타적이고, 어느 대상이든 한 그룹에만 소속되어야 한다.

㉡ 분류가 논리상 정연해야 하며

㉢ 가능한 한 각 그룹에 속하는 표본의 수가 같거나, 비슷한 것이 좋다.

② 독립변수의 선택 ⇒ 기존 연구나 경험으로부터 얻으며, 종속변수의 그룹 별로 차이를 나타낼 수 있는 변수들을 선택하는 것이 좋다.

2) 가정의 점검

① 각 독립 변수들의 분포가 정규분포인가?
→ 결합확률분포가 다변량 정규분포를 하는 가를 예측.
② 그룹 별 독립 변수들은 등분산성을 만족하는가?
→ Box-M 검정을 이용
③ 각 그룹에 소속 될 사전확률은 같은가?

3) 표본의 크기
① 전체 표본의 크기는 독립변수의 개수보다 3배(최소한 2배) 이상 되어야 한다.
② 종속변수의 그룹 각각의 표본의 크기 중 제일 작은 것이 독립변수의 개수보다 커야 한다.

4) 판별함수에 포함될 독립변수의 선택방법 ⇒ 회귀분석과 같음.
① 동시 포함법(=직접 선택법)
② 단계적 추출법→ stepwise 법
③ 후진 제거법
④ 전진 선택법

5) 독립변수의 선정 기준

⇒ 판별함수에 포함될 독립변수를 단계적으로 선정할 때 (즉, stepwise 법, 전진선택이나 후진제거법), 한 변수씩 추가하거나 제거하고자 할 때 고려 대상이 되는 기준.

① 윌크스 람다(Wilks` Lambda : Λ)

⇒ Λ 값이 최소가 되는 변수를 선택함.

$$\because \quad \Lambda = \frac{\text{그룹 내 분산}}{\text{총 분산}}$$

→ 만약 Λ값이 커지면, 그룹 사이의 분산보다 그룹 내 분산이 크다는 의미가 되어서 그룹 사이의 차이를 바로 설명하지 못함.

② Maholanobis 거리 : D-statistic
⇒ 그룹이 2개 일 경우에만 사용

$$D^2 = (\mu_1 - \mu_2) W^{-1} (\mu_1 - \mu_2)^T$$

여기서, μ_1 , μ_2 : 각각 그룹 1과 2의 평균 벡터
W : 그룹 1과 2의 공분산 행렬

→ D^2 값이 가장 큰 변수를 택함.

③ 직접 지정법 ⇒ 앞의 방법들을 이용하여 최종 모형을 확정한 다음, 꼭 필요한 변수들을 추가 한다.

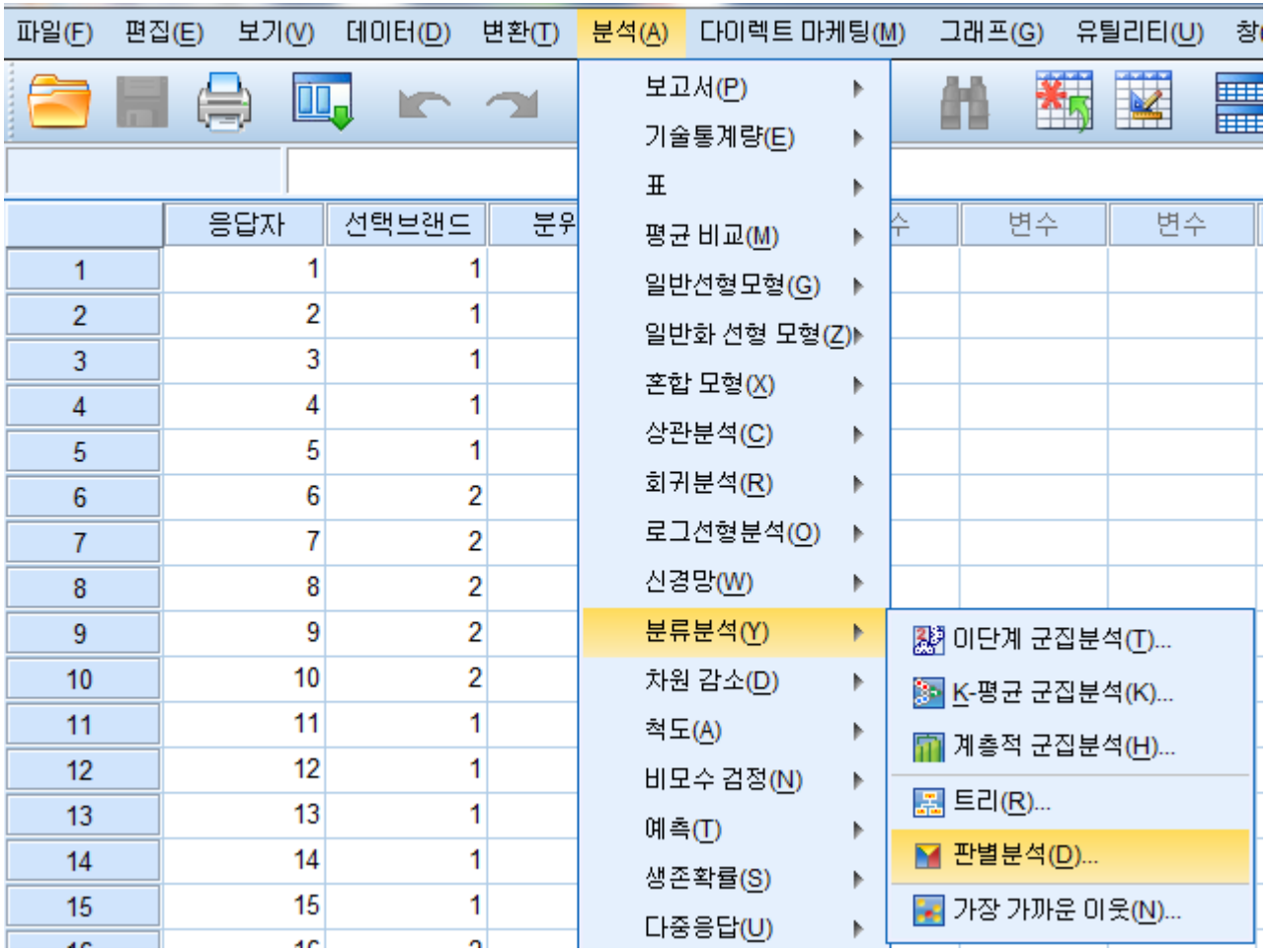
2. 가설검정을 위한 판별분석

2.1-2. 분석개요 및 데이터

- 분석개요
  - 판별분석을 위해 설명변수로 선택된 변수들이 판별집단에 해당되는 종속변수에 대해 영향을 미치는 정도를 보기 위한 것
- 분석데이터(p.415) <10장-2-1-1.데이터.sav>
  - A회사에서 외국브랜드를 선호하는 소비자와 국내브랜드를 선호하는 소비자의 특성을 파악하기 위하여 10명의 외국브랜드 선호자와 10명의 국산브랜드 선호자의 구매를 조사하여 구매시 디자인과 가격의 중요성을 11점 척도로 측정한 데이터

■ SPSS를 이용한 분석

▷ 1단계 ◁ <분류분석(Y)> → <판별분석(D)…>을 선택

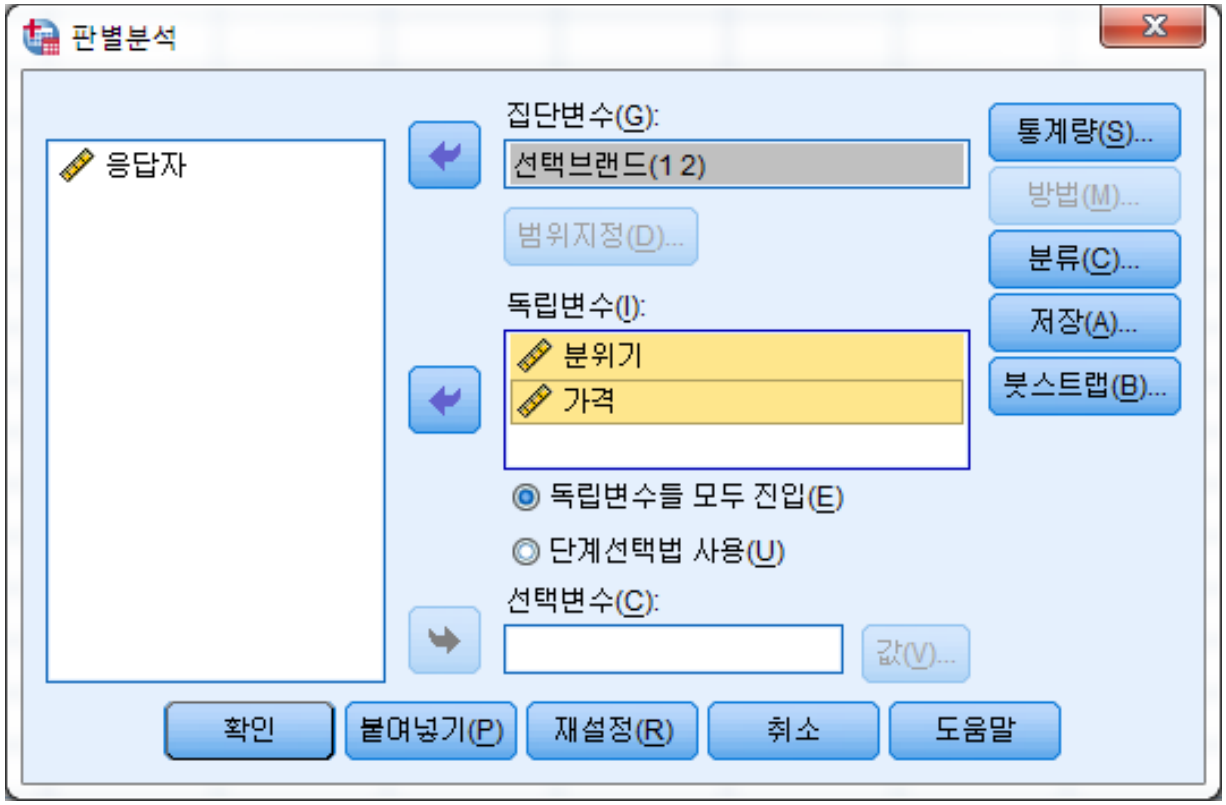


▷ 2단계 ◁ <판별분석> 대화상자에서

① 집단변수 : '선택브랜드'를 선택하고, 범위를 1~2까지 지정

② 독립변수 : 분위기, 가격을 지정

③ 변수선택법 : '독립변수를 모두진입' 지정

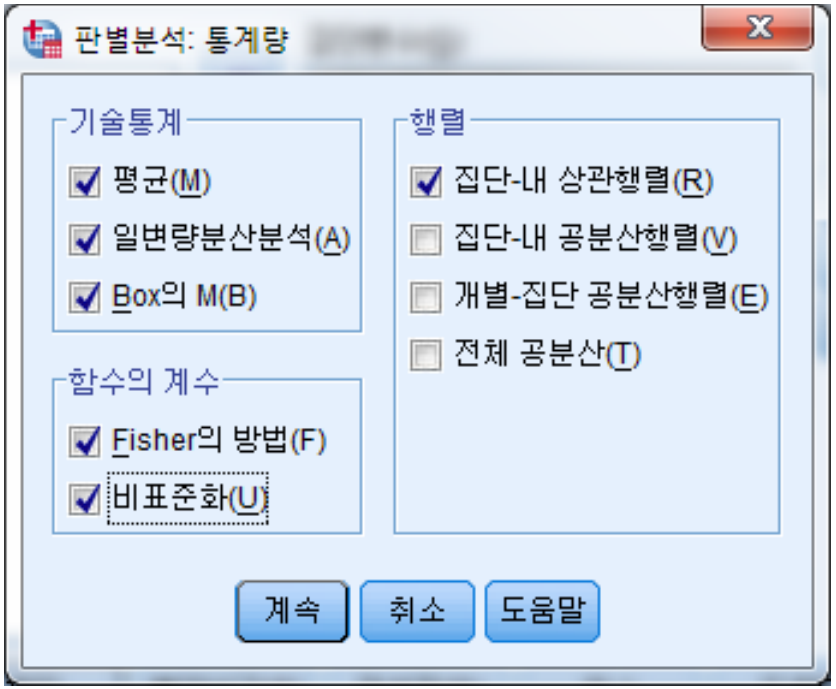


▷ 3단계 ◁ <통계량(S)…>을 click

① 그룹 별 독립 변수들은 등분산성을 만족하는가?
→ Box-M 검정을 이용

② 비표준화 방법을 이용하여 판별함수를 계산
→ 비표준화 분류계수를 이용하여 판별점수를 계산하고 분류.

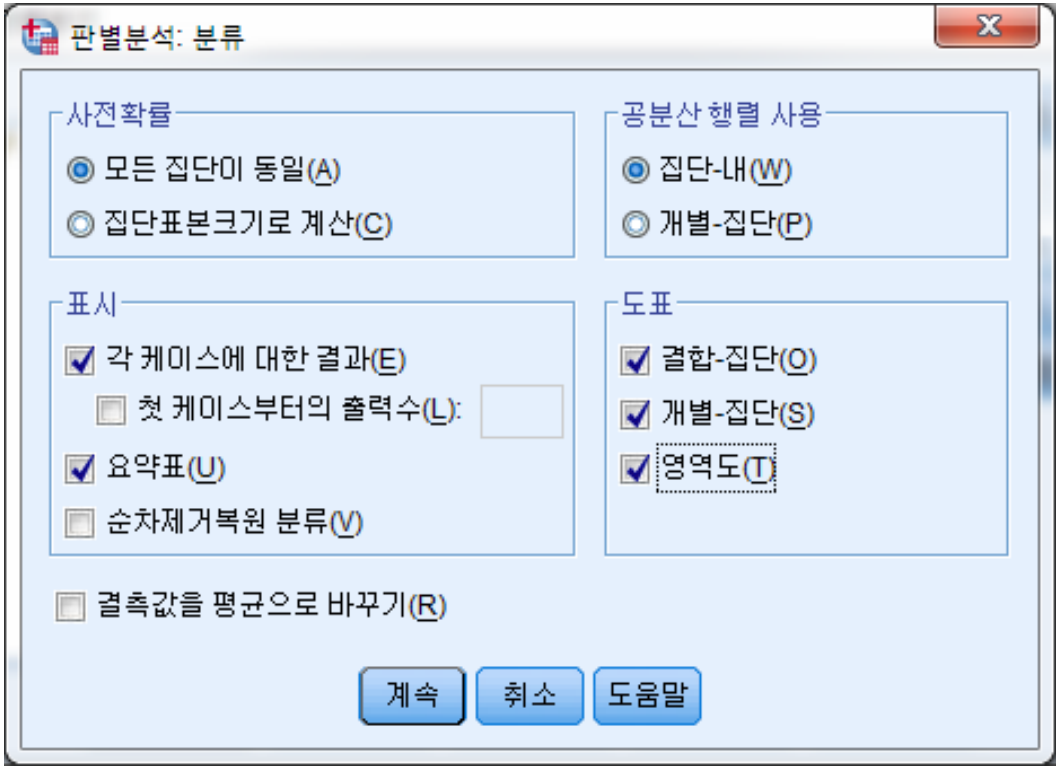
③ 변수들 사이의 상관관계의 정도를 알아보기 위해
→ 집단 내 상관행렬을 구함.



▷ 4단계 ◁ <분류(C)…> click

① 사전확률은 모든 집단이 동일하게

② 필요한 결과 및 도표를 출력



▷ 5단계 ◁ <저장(A)>을 click

① 판별분석 결과 예측소속 집단을 알아 봄

② 판별점수 및 집단 소속 확률을 click