CHAPTER 1

다변량 분석이란

다변량 분석은 1)변수들 간의 인과 관계(casual relationship)를 규명, 분석하거나 (다중 회귀 분석: Multiple Regression, 다변량 분산 분석: Multivariate ANOVA) 2)변수들 간의 상관 관계를 이용하여 변수를 축약(reduction)하거나 개체들을 분류(classification)하는데 관련된 분석 방법이다. 일반적으로 협의의 다변량 분석이란 후자를 일컬으며 변수 개수와 개체 개수가 많은 대용량(large), 복잡(complicate & complex) 자료에 대한 분석 방법이다. 이 책에서는 이 부분만을 다룰 것이다. 다변량 분석의 예를 들어 보자.

(1)변수 축약

학생들의 지적 수준을 측정하기 위하여 수학, 영어, 국어, 과학 4 개 과목 수학 능력을 측정하였다고 하자. 학생들의 4 과목 성적을 살펴보았더니 수학이 높은(낮은) 학생은 과학 성적이 높고(낮고), 영어가 높은(낮은) 학생은 국어 성적이 높음(낮음)을 알았다. 이 경우 수학과 과학 성적이 공통적으로 측정하는 공통 개념(common entity)이다. 이를 수리 능력이라 하자. 또한 영어와 국어 성적이 공통적으로 측정하는 공통 개념(예를 들면 언어 능력)이 있을 것이라는 것을 알게 될 것이다. 이처럼 변수들간의 상관 관계를 이용하여 4 개 변수(수학, 영어, 국어, 과학)를 2 개의 새로운 변수 집단(수리, 언어)으로 그룹화할 수 있다. 이를 요인 분석이라 한다. 4 개의 원 변수의 선형 결합에 의해 만들어진 주성분을 이용하여 변수의 차수를 줄이는 방법이 주성분 분석이다.

(2)개체 분류

통계학과 재학생 100 명에 대해 IQ, 평점, 키, 몸무게, 성별, 가계소득, 용돈 등을 측정하였다고 가정하자. 측정 변수들 간의 상관 관계를 이용하여 100 명을 유사한 몇 개의 그룹으로 나눌 수 있을 것이다. 이를 군집 분석이라 한다. 다변량 데이터에 집단 분류 변수가 있다면 개체의 집단을 판별하는 판별식을 구하고 새로운 개체를 분류하는 방법으로 판별 분석도 있다.

(3)변수들의 선형 결합

신체적 조건(키, 몸무게, 가슴둘레)과 운동력(달리기, 윗몸 일으키기, 턱걸이) 사이의 선형 상관 관계, 즉 변수들 간의 선형 결합 관계를 분석하는데 다변량 분석이 이용된다. 이에 적합한 분석 방법은 정준 상관 분석이다.

통계적 연구 방법은 데이터를 수집하고 적절한 통계적 연구 가설 설정, 적합한 분석 방법을 이용하여 연구 가설 채택 여부를 알아보는 확증적(confirmatory) 연구와 수집된 데이터를 정리, 분석하여 자료에 포함된 정보를 얻어내는 탐색적((exploratory) 연구가 있다. 이러한 연구 방법에서 볼 때 다변량 분석은 확증적인 분석 방법이 라기 보다는 탐색적 분석 방법에 가깝다.

1.1. 데이터

1.1.1. 데이터 행렬

통계학은 정보를 가진 숫자들의 모임인 데이터에 관한 학문(Statistics is about data) 이다. 데이터 수집(collection), 정리(summarization), 분석(analysis), 표현(presentation)을 통해 데이터가 가진 정보를 얻어내는 과정에 관한 학문이 통계학이다.

통계학에서 데이터라 함은 관심 대상이 되는 집단으로부터(모집단: population) 추출한 표본(sample) 개체의 (예: 사람, 동식물, 기업) 특성치(변수, 예: IQ, 체중, 바이러스 수, 불량 여부, 도매 물가 지수)를 말한다. 즉 데이터는 변수와 관측치로 구성되어 있다.

데이터를 수집하는 방법은 기존의 데이터를 수집하는 방법(예: 경제 관련 지수, 기업 경영 재무제표 관련 특성치), 실험에 의한 방법(예: 신약의 효과 검증), 측정 및 관측 방법에 의한 방법(예: IQ, 비만도 지수, 교통량), 설문에 의한 방법(예: ○○ 후보 지지율, △△ 법안

찬성 여부) 등으로 나눌 수 있다. 때로는 실험에 의한 방법과 비 실험 방법으로 나누기도 한다.

데이터를 수집한다는 것은 모집단 특성 중 (1)관심이 있고 (2)측정 가능한 것을 표본 개체를 통해 얻는 것을 의미하며 각 특성을 변수라 하고 측정된 각 표본 개체의 특성 값을 관측치라 한다. ○○ 대학생들의 지적 능력 가운데 수리 능력과 언어 능력에 관심이 있다고 하자. 이 경우 수리 능력은 교양 수학 성적으로 영어 능력은 토익 성적으로 측정할 수 있을 것이다. 학생들 30 명을 확률 표본으로 추출하여 수학 성적과 토익 성적을 조사하는 것을 데이터 수집이라 한다. 수학 성적, 토익 성적이 변수이고 조사된 각 학생들의 수학 성적, 영어 성적은 관측치이다.

데이터를 정리하거나 통계 소프트웨어의 사용을 위해 코딩 할 때는 열은 변수, 행은 개체로 하여 행렬의 형태로 나타내는데 이를 데이터 행렬(data matrix)이라 한다. 위의 예제 자료를 행렬의 형태로 표현하면 다음과 같다. 예를 들어 n 명의 사람에 대해 p 개의 변수(성별, 키, 몸무게, 학력, 재산 정도, …… 결혼 여부)를 측정하였다고 하자. 이 때 데이터 행렬은 다음과 같다. 자료 행렬의 각 행을 자료 벡터(타원)라 하고 열은 변수 벡터(직사각형)라 한다. 다음은 실제 데이터와 데이터 행렬의 원소 x_{ij} 를 대응시켜 행렬에 대한 이해를 돕기 위해 작성한 표이다.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

		var_1	var_2	var_3	…	var_p
개체1	obs_1	남자	180	82	…	Married
		X_{11}	X_{12}	X_{13}		X_{1P}
개체2	obs_2	여자	163	56	…	Single
		X_{21}	X_{22}	X_{23}		X_{2P}
…	…	…	…	…	…	…
개체 n	obs_n	남자	173	75	…	Single
		X_{n1}	X_{n2}	X_{n3}		X_{np}

다변량 분석에서 변수만을 표시할 때 변수 벡터는 다음과 같이 표현한다. 변수 벡터의 x_i 들은 변수를 의미하므로 p는 변수의 개수이다.

$$\underset{\sim}{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \quad \text{혹은} \quad \underset{\sim}{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}, j = 1,2,\cdots,n$$

1.1.2. 데이터의 역사

데이터는 원하는 정보를 얻기 위하여 분석자가 분석 목적에 맞는 변수를 설정하고 적절한 자료 수집 방법, 표본 추출 방법을 설계하여 수집한 변수와 관측치의 모임이다. 그러므로 누가, 언제, 어디서, 어떤 수집 방법과 표본 추출 방법을 이용하여 데이터를 수집되었는지에 대한 정보가 있어야 올바른 데이터 분석이 가능하다. 이것을 데이터의 역사(history) 혹은 흔적(trail)이라 한다. 데이터는 정보를 가진 숫자의 모임으로 그 숫자를 얻는 과정, 즉 데이터의 역사를 알아야만 올바른 분석을 설정하고 정보의 가치를 판단할 수 있다.

데이터 역사는 데이터 분석 결과를 해석하는데 중요하다면 변수의 형태는 적절한 분석 방법을 결정하는데 판단 기준이 된다.

1.1.3. 변수의 종류

수리 통계에서 (확률)변수는 확률 실험(random experiment, 실험 결과를 예측할 수 없는 실험)에서 발생 가능한 모든 결과의 모임인 표본 공간(sample space, S)의 원소(element)와 실수 값을 대응시킨 함수라 정의된다. 데이터 분석에서의 변수란 분석자가 관심을 갖는 각 개체의 특성을 말한다. 변수의 의미는 개체에 따라 특성 값이 변한다는 의미이다.

▣수리 통계 측면

(1)연속형(continuous)

실험 결과가 무한히 많은 변수이다. 아무리 작은 임의의 어떤 임의의 구간을 택하더라도 그 구간 내의 하나 이상의 값이 측정 가능한 경우로 키, 몸무게, 소득 등이 여기에 해당된다.

(2)이산형(discrete)

측정 결과를 셀 수 있는 경우로 성별, 직업, 교통량, 나이 등이 여기에 해당한다.

▣자료 분석 측면

(1)측정형 변수(metric, measurable, quantitative)

셀 수 있거나 측정 가능한 특성을 측정한 변수로 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측정형 변수이고 이산형 변수 중 측정형 변수가 있을 수 있다. (예: 교통 사고 건수)

(2)분류형, 범주형 변수(Non-metric, categorical, classified)

개체를 분류하기 위해 측정된 변수를 의미하며 성별, 결혼여부, 학년, 소득(상, 중, 하) 등이 그 예이다. 분류형 변수는 범주의 형태에 따라 다음과 같이 분류된다.

①명목형(nominal)

범주의 크기 순서가 없는 경우로 성별(남, 여), 결혼 여부(미혼, 기혼)가 이에 속한다.

②순서형(ordinal)

범주의 크기 순서가 있는 경우로 성적(A>B>C>D>E)이 그 예이다.

▣시간에 의해

자료가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고 그렇지 않은 경우를 횡단면 자료(Cross-section: 일정 시간에 한꺼번에 조사)라 한다. 경제 지표(환율, 수출량)나 기업의 연별 자료(연도별 매출액)가 시계열 자료에 해당된다.

▣인과 관계 속에서

인과 관계(casual relationship)에서 원인이 되는(영향을 주는) 변수를 설명(exploratory) 변수 혹은 독립(independent) 변수라 하고 결과(영향을 받는 변수)를 종속 변수(dependent) 혹은 반응(response) 변수라 한다. 종속 변수는 Y, 설명 변수는 X로 표시

한다. 분산 분석에서는 설명 변수를 처리 효과 혹은 요인으로 불리어진다. 인과 관계는 이론적, 경험적 타당성에 근거하여 연구 목적에 설정되는 것이지 자료 분석 후 인과 관계가 설정되는 것은 아니다.

협의 다변량 분석은 인과 관계에 대한 분석이 아니라 변수들간의 상관 관계를 이용하므로 종속 변수는 존재하지 않는다.

1.2. 다변량 분석의 종류

다변량 분석을 정리하면 다음과 같다. 협의의 다변량 분석은 1.2.2 절의 분석 방법을 의미하며 본 책은 이 방법들만을 다룰 것이다.

1.2.1. 종속 변수와 독립 변수 사이의 인과 관계

(1)다중 회귀(Multiple Regression)

종속 변수는 측정형 변수, 설명 변수들은 모두 측정형 혹은 측정형과 분류형(이를 지시 변수 혹은 가 변수라 한다.)이 혼합되어 있는 경우로 종속 변수가 2 개 이상인 경우를 다변량 회귀분석(Multivariate Regression), 회귀 모형들간의 상호 관계를 분석하는 경우를 연립 방정식 회귀(Simultaneous Equation Regression) 분석이라 한다.

||예제||

사람들의 재산 정도(단위: 천원)에 영향을 미치는 변인으로 교육 정도, 부모의 재산 정도, 나이, 연봉을 생각해보자. 회귀 분석은 1)재산 정도에 교육 정도, 부모의 재산 정도, 나이, 연봉이 영향을 미치는지(회귀 계수의 유의성) 2)영향을 미친다면 어떤 영향(회귀 계수의 부호)으로 얼마나(회귀 계수의 크기) 영향을 미치는지 3)어떤 변수의 영향력(표준화 회귀 계수)이 가장 큰지를 알아볼 때 사용된다.

(2)로지스틱 회귀(Logistic Regression)

종속변수가 이진(binary, dichotomous) 변수이거나 순서형 변수인 경우 사용되는 회귀 분석 방법이다. 로지스틱 회귀 분석은 개체 판별에도 이용된다.

||예제||

○○대학 졸업 예정자들의 취업 여부에 졸업 학점, 토익 성적, 외국 어학 연수 기간(단위: 월), 전공 과목 선택 비율 등이 영향을 미치는지 알아볼 대 로지스틱 회귀 분석이 사용된다.

(3)분산 분석(ANOVA: Analysis Of Variance)

종속변수가 측정형이고 설명변수가 모두 분류형인 경우 분산 분석이 사용된다.

||예제||

재산 정도(단위: 천원)에 교육 정도(고졸 이하, 대졸, 대학원졸), 부모의 재산 정도(상, 중, 하), 나이(30 대, 40 대, 50 대)에 다른 차이가 있는지 알아보고자 할 때 분산분석이 사용된다.

(4)다변량 분산분석(Multivariate ANOVA)

측정형 변수인 종속변수가 2 개 이상인 분산 분석으로 결과 해석이 용이하지 않아 많이 쓰이지 않는 분석 방법이다.

||예제||

사람의 근심지수와 불면지수(종속변수)에 성별, 교육 정도, 재산 정도(상, 중, 하), 가족 구성 종류(핵가족/대가족)가 영향을 미치는지 알고자 할 때 다변량 분산 분석이 사용된다.

1.2.2. 변수 축약

변수들간의 상관 관계를 이용하여 변수를 줄이는 방법으로 변수 유도 기법(variable directed technique)이라고도 한다. 변수 유도 기법들은 원 변수들로부터 새로운 변수를 만들게 되는데 이는 성분(component), 요인(factor), 정준 변수(canonical variate) 등으로 불린다.

(1)주성분 분석(Principal Component Analysis: PCA)

(2)요인 분석(Factor Analysis: FA)

(3)정준 상관 분석(Canonical Analysis: CA)

정준 상관 분석은 변수 군간의 상관 관계를 살펴보는 방법이므로 엄밀히 말하면 변수 축약 방법은 아니다.

1.2.3. 개체 분류

개체들의 특성을 측정한 변수들의 상관 관계를 이용하여 유사한 개체를 분류하는 방법으로 개체 유도(Individual Directed) 기법이라 한다.

(1)군집 분석(Cluster Analysis: CA): Multi-Dimensional Scaling (MDS: 다차원 척도법)

(2)판별 분석(Discriminant Analysis: DA): 정준 판별 분석(Canonical Discriminant Analysis: CDA), 로지스틱 판별 분석 (Logistic Discriminant Analysis: LDA)

변수 유도 기법이나 개체 유도 기법은 변수들간의 상관 관계로부터 시작되므로 모든 변수들이 측정형이거나 적어도 순서형 변수이어야 한다.

1.3. 다변량 분석 맛보기

본 책에서 다루게 될 협의의 다변량 분석을 간단히 살펴보자. 예제로 ○○대학 학생 30 명(장학생 15 명, 일반 학생 15 명)을 수학, 과학, 영어, 국어 능력을 측정했다고 가정하자.

1.3.1. 주성분 분석

다변량 자료에 존재하는 비정규성(abnormality)이나 이상치(outliers: 측정된 변수들 면에서 다른 개체들과 상이한 개체)를 발견하기 위하여 변수들의 상관 관계(공분산 행렬 혹은 상관 행렬을 이용하여) 상관 관계가 존재하지 않는 새로운 변수(주성분)를 구한다. 주성분 분석은 개체들을 순서화 하거나 분류하는데 사용되기도 하고 회귀 분석 시 발생하는 다중 공선성 문제 해결 방법으로 이용된다. 이처럼 주성분 분석은 최종 분석이라기 보다 초기 분석이라 보는 것이 더 적절할 것이다. 주성분 분석은 p 개 변수들로부터 서로 독립인 $k(< p)$ 개 주성분을 구해 원 변수의 차원을 줄이는 방법이다.

학생들의 수학, 과학, 영어, 국어 성적의 상관 관계를 이용하여 서로 독립인 2~3 개의 주성분 변수를 얻게 되는데 주성분 변수는 원 변수의 선형 결합에 의해서 얻어진다.

1.3.2. 요인 분석

측정 변수들을 그룹화 하는데 사용된다. 원 변수(수학, 과학, 영어, 국어)를 설명하는 내재 변인(이를 요인이라 한다.)에 의해 원 변수를 그룹화 하는 분석 방법이다. 예를 들어 수리 능력 변인에 의해 수학, 과학능력이 잘 설명되면(►요인 분석에서 자세히 설명하기로 한다.) 수학, 과학을 평균 내어 수리능력 점수로 이용하게 된다. 주성분 분석과 요인 분석은 변수들 간의 관계를 새로운 변수(이름은 다르지만)를 이용하여 살펴본다는 면에서 보면 유사하나(variable-directed technique) 요인 분석은 변수들을 그룹화하는 것이 목표이며 주성분 분석은 새로운 변수인 주성분을 이용하여 원 변수의 차수를 줄이는 것이 주 목적이다. 요인 분석은 설문 조사에서 가장 많이 사용되는데 동일 개념을 측정한 리커드 척도 문항들을 분류할 때 사용된다.

1.3.3. 판별 분석

이미 2 개 이상의 그룹으로 나누어진 개체들에 대해 분류에 영향을 미칠 것 같은 특성(변수)을 측정하고 이를 이용하여 판별식을 구해 새로운 개체를 분류하는 방법이다. 예를 들어, 신용 카드 회사에서 신용도를 평가하여 고객들을 우량, 보통, 불량으로 구분하고 신용 판별에 적합하다고 생각되는 변수(성별, 재산 정도, 월 수입, 학력 등)를 조사하였다고 가정하자. 이 변수들을 이용하여 신용도 평가 기준(판별식)을 설정하고 새로운 신청자가 오면 성별, 재산 정도, 월 수입, 학력 등을 조사 후 판별식에 의해 신용도를 평가하게 된다. 판별식을 Logistic 회귀 분석을 이용하여 구하는 경우 이를 Logistic 판별 분석이라 한다.

학생 30 명의 장학금 수혜 여부를 판단할 수 있는 판별식을 원 변수(수학, 과학, 영어, 국어)에 의하여 유도하고 새로운 학생이 들어오면 수학, 과학, 영어, 국어 능력을 측정하여 그 판별식에 의해 장학금 수혜 가능성(확률)을 판단하게 된다.

1.3.4. 군집 분석

군집 분석은 인류학자가 발굴 과정에서 발견한 암석들에 대해 화학 성분들을 측정하여 암석들을 분류하고자 할 때 사용되는 분석 방법이다. 군집 분석과 판별 분석은 개체들을 그룹화 한다는 면에서는 유사하나 판별 분석은 분석 이전에 개체들의 그룹이 정해져 있고 군집 분석은 분석을 통해 적절한 그룹 개수가 결정되고 개체가 분류된다. 개체들의 특성을 측정한 변수들을 이용하여 개체들간의 유사성을 측정하고 이를 이용하여 개체들을 저 차원 가시적 공간(2 차원)에 표현하는 방법인 다차원 척도법(MDS: Multi-Dimensional Scaling)도 군집 분석의 일종이다.

군집 분석은 개체 들의 분류가 주 목적이고, 판별 분석은 새로운 개체의 분류가 주 목적이다. 판별 분석과는 달리 30 명 학생의 군집(그룹, 여기서는 장학금 수혜 여부)에 대한 정보가 없다. 수학, 과학, 영어, 국어 성적을 이용하여 학생 30 명을 적절히 분류하게 된다. 군집의 개수는 분석자가 임의로 정한다.

1.3.5. 다변량 분산 분석

다변량 분산 분석은 종속변수가 2 개 이상인 분산 분석 방법이다. 예를 들어 근심 지수, 불면지수 그리고 불쾌 지수(종속 변수 3 개)에 성별, 학벌, 재산 정도(상, 중, 하), 가족 구성 종류(핵가족, 대가족)가 영향을 미치는가를 동시에 분석하고자 할 때 사용된다. 근심 지수, 불면 지수, 불쾌 지수를 각각 종속변수로 하여 3 개의 분산 분석(ANOVA)을 개별적으로 실시할 수 있으나 다음과 같은 문제가 발생한다.

(1)근심 지수, 불만 지수, 불쾌 지수는 성별(설명 변수, 요인)에 따라 다를 수 있고 종속변수들간 상관 관계가 존재하므로 종속 변수와 설명 변수를 동시에 고려한 분산 분석이 필요하다.

(2)1 종 오류 문제이다. k 개의 분산 분석을 개별적으로 실시하면 1 종 오류(유의 수준)는 더 이상 α 가 아니다. 유의 수준이란 귀무가설이 참인데도 이를 기각할 오류(확률)이므로 $1-(1-\alpha)^k$ 이다. 그러므로 더 이상 우리가 설정한 유의수준은 유효하지 않다. 다변량 분산 분석에 의해 요인의 효과가 존재한다고 분석되면 어떤 종속 변수에서 요인 효과가 있는가 개별적 분산 분석을 실시해도 되나 그렇지 않은 경우 “거짓 유의성(false significant)”이 발생하므로 분석 결과를 해석하는데 어려움이 있어 비록 1 종 오류 문제가 있지만 종속변수 각각 분산 분석을 실시하는 경우가 빈번히 발생한다.

1.3.6. 정준 상관 분석

예를 들면 부모의 양육 태도에 관련된 변수 군과 딸의 연애 행동에 관련된 변수 군의 상관관계를 보는 것으로 다중 상관 계수, 편 상관 계수와는 구별 된다. 각 변수 군의 변수들이 선형 결합에 의해 새로운 변수가 만들어지게 되는데 이를 정준 변수라 하고 정준 변수의 상관 계수를 구하므로 정준 상관 분석이라 한다.

다변량 분석 방법을 정리하면 다음 표와 같다.

	주성분 분석	요인 분석	판별 분석	군집 분석	다변량 분산 분석	정준 상관 분석
변수들 관계 탐색	S	D	N	N	N	S
자료 탐색	D	S	N	S	N	N
새 변수 만들기	Yes	Yes	No	No	No	Yes
개체 분류	No	No	Yes	Yes	No	No
그룹간 평균 비교	P	P	R	R	Yes	No
변수 그룹	P	P	N	N	N	D
차원 줄이기	D	P	N	N	N	N

N: 불가능, P: 가능, R: 드묾, S: 때때로, D: 언제나