계량분석

Construct Validity and Principal Component Analysis

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

November 28, 2024

진행 순서

1. 타당도
2. 요인분석과 주성분분석
3. 주성분분석의 원리
4. 두 개 이상의 주성분

타당도

타당도

연구가 "어디에서" 타당한가에 관해서 중요한 쟁점이 있다.

- 내적 타당도(internal validity)는 연구에서 제기된 인과관계가 보다 엄밀하게 성립함을 보였을 때 확보된다. 근본적으로 실험자료(experimental data)가 아닌 관찰자료(observational data)를 사용했을 때는 관찰되지 않은 이질성(unobserved heterogeneity)을 제거할 수 없으므로 내적 타당도는 필연적으로 훼손된다.
- 외적 타당도(external validity)는 실험실 바깥의 대상에 대해서도 얼마만큼 연구 결과가 일반화 될 수 있는가에 따라 담보된다. 많은 임상시험(clinical trials)의 피실험자 집단은 백인(보다 구체적으로 남자-대학 2학년)으로 대표되며 대체로 다양성을 고려하지 않으므로 외적 타당도가 훼손된다(는 비판이 있다).
- 일반적으로 실험설계(design of experiment; DoE)는 내적 타당도가, 서베이설계는 외적 타당도가 높다. 다만 이 주제는 사실 이렇게 대충 넘어갈 수 없고 생각보다 복잡한 이슈들을 안고 있다.

타당도

확보되어야 하는 요건에 따라서도 타당도의 종류가 구별된다.

- 우선 (1) 구성 타당도(construct validity)로 사회통계학의 맥락에서는 다른 어떤 것보다 먼저 논의된다. “이론적 예측대로” 다른 지표/변수들이 관계 구조를 가지면 구성 타당도가 높다고 할 수 있다.
- 이른바 요인분석(factor analysis)을 통해 구성 타당도를 탐색 또는 확인한다.
- 높은 구성 타당도는 두 원리의 충족을 요구하는데, 그것은 “같은 것은 같게 (convergent), 다른 것은 다르게(discriminant)”이다.
- 전자를 수렴 타당도(convergent validity)라고 부르고, 후자를 판별 타당도 (discriminant validity)라고 부른다.

타당도

- 만약 그 다음으로 중요한 부분이 있다면 그것은 (2) 예측 타당도(predictive validity)이다. 이것은 지표 점수가 기대했던 미래의 상태와 높은 연관성을 가지면 확보된다.
- 교과서/수험서에 따라서 이는 동시 타당도(concurrent validity)같은 개념과 결합하여 기준관련 타당도(criterion-related validity)라고 불리기도 한다. 동시 타당도는 기존의 타당화된 검사/지표와 비교함으로써 확보한다.
- 최근 데이터과학의 진보와 더불어 점차 그 중요성을 더해가고 있다(고 나는 생각한다).

타당도

- (3) 내용 타당도(content validity)는 측정항목(measurement items)이 충분히 포괄적(inclusive)인가를 따진다.
- 만일 사칙연산 능력을 측정하려는데 나누기를 확인하는 항목이 빠져있다면 내용 타당도가 결여된 셈이다.
- 연구자의 전문성에 의거하여 확보되는 (4) 액면 타당도(face validity)도 있다.
- 보고서 등에서 종종 "자문을 구하고 포커스그룹 인터뷰(FGI) 등을 수행하여 액면 타당도를 확보하고자 노력하였다"는 식의 언급을 볼 수 있다.
- 교과서/수험서에 따라서는 내용 타당도와 액면 타당도를 같은 것으로 묶는다.

타당도

이론적 구성물은 타당성을 여러 측면에서 확보해야 한다.

- 복합적인 사회현상에 관한 이론적 구성물을 추론하기 위해 단 하나의 측정문항만으로는 충분히 포괄적이지 않다. 그러므로 여러 개의 측정문항들을 사용하여 개념을 타당성 있게 측정한다(내용 타당도).
- 그런데 복수의 측정문항들은 측정하고자 하는 개념을 중심으로 응집력있게 모여야 한다(수렴 타당도).
- 뿐만 아니라 이 측정문항들은 (설령 다소 비슷하더라도) 다른 개념을 위한 측정문항으로부터 충분히 구별될 수도 있어야 한다(판별 타당도).

타당도

- 예를 들어 “또래집단 지지”와 “가족 지지”라는 두 개념이 있을 때,

  (1) 또래집단 지지를 측정하고자 하는 문항들은 또래집단 지지를 중심으로 충분히 높은 수렴 타당도를 가져야 하고,

  (2) 가족 지지를 측정하고자 하는 문항들에 대해서도 충분히 높은 판별 타당도를 가져야 한다.
- 이와 같은 두 차원(또래집단 지지와 가족 지지)은 더 상위 개념의 일부(예컨대 사회적 지지)가 될 수도 있다.

요인분석과 주성분분석

요인분석과 주성분분석

이른바 요인분석을 통해 구성 타당도를 계량적으로 평가할 수 있다.

- 광의의 요인분석은 공통요인분석(common factor analysis)과 주성분분석(principal component analysis)을 포괄한다.
- 협의의 요인분석은 보통 공통요인분석만을 뜻한다.
- 기계학습(machine learning) 관점에서 요인분석과 주성분분석 모두 차원 축소(dimensionality reduction)를 수행한다.
- 사회과학연구의 맥락에서 차원 축소란 변수가 너무 많을 때 이를 압축하여 몇 개의 변수로 줄여나가는 것을 뜻한다.

요인분석과 주성분분석

- 다만 요인분석은 데이터 속에 어떤 공통된 요인(common factor) 내지 잠재된 구조가 있다고 전제하고 이를 밝혀내는데 관심을 갖는다.
- 그러므로 훨씬 심리학 혹은 사회과학의 측면에서 이론지향적 성격을 갖는다.
- 반면 주성분분석은 정보의 손실을 최소화하면서 차원의 저주(curse of dimensionality)를 푸는데 더 직접적인 관심을 갖는다.
- 수학적으로 주성분분석이 요인분석보다 간결하고 아름답다(고 나는 생각한다).

요인분석과 주성분분석

- 둘 다 변수들 사이의 상관계수행렬(correlation coefficient matrix)을 분석의 원료로 사용한다.
- 그런데 주성분분석은 대각행렬 부분에서 1을 그대로 사용하는 반면, 요인분석은 이 값을 공통분(communality) 값으로 대체한다(공통분이 무엇인가는 이제 곧 배우게 된다).
- 이로 인해 주성분분석은 변수의 공통분산(common variance)과 고유분산(unique variance), 오차분산(error variance)을 가리지 않고 모두 사용하는 반면, 요인분석은 엄격하게 공통분산(common variance) 부분만을 사용한다(고 심리학자와 교육학자들은 특히 강조한다).

요인분석과 주성분분석

- 이 수업에서는 주성분분석의 기초와 응용만을 간략히 학습하고 요인분석은 전혀 다루지 않는다. 요인분석은 그 자체로 최소 한 학기의 수업이 필요하다(기회가 닿으면 꼭 듣자).
- 특히 확인적 요인분석(confirmatory factor analysis)은 이른바 구조방정식모형(structural equation modeling) 학습의 출발점이 되고, 탐색적 요인분석(exploratory factor analysis)은 척도 개발(scale development)에 필수불가결하게 쓰이는 기법이다.
- 척도 개발(e.g., 여성혐오 척도)은 그 자체로 굉장히 큰 방법론적 분야이므로 설명이 쉽지 않다.

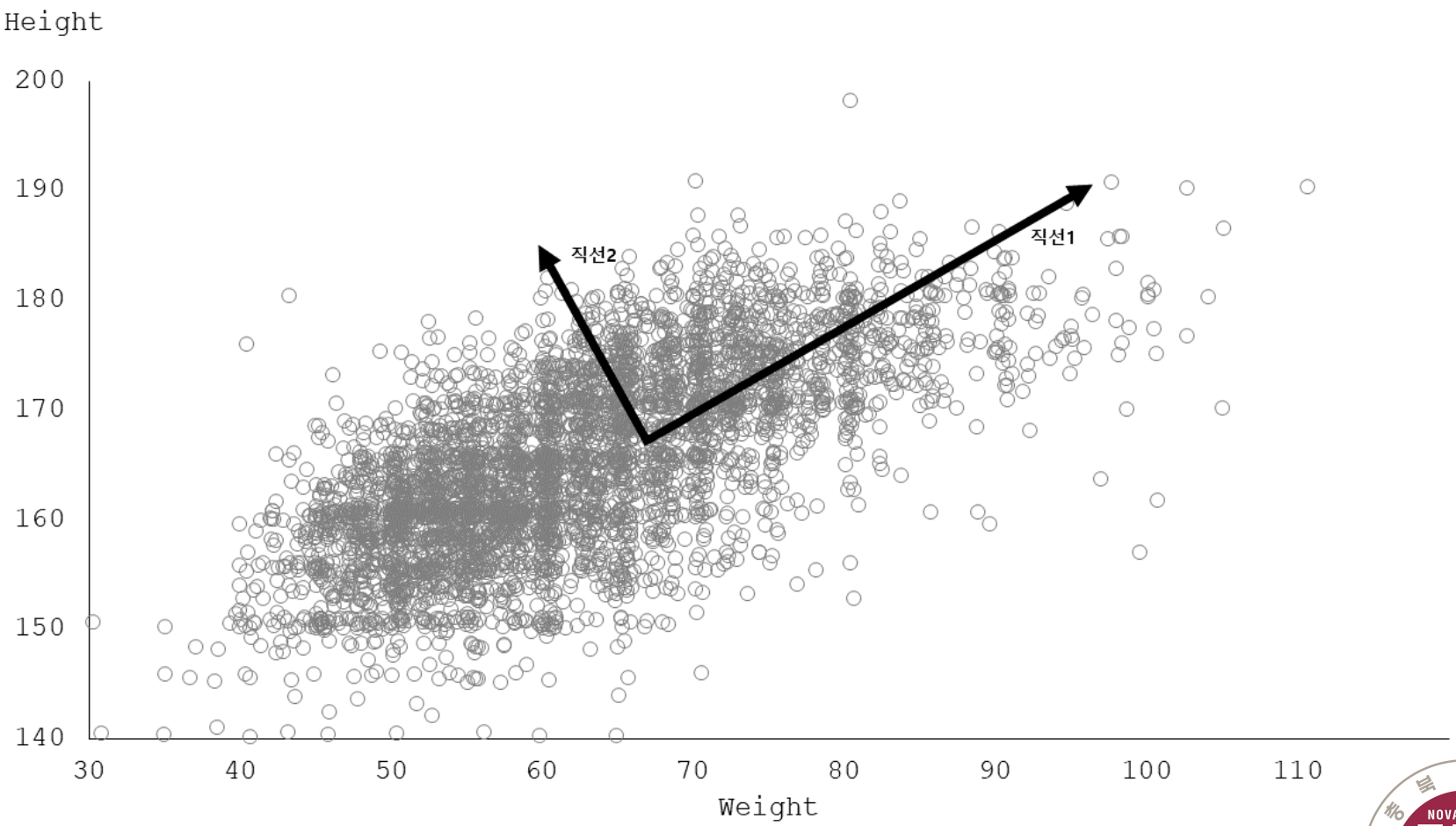
주성분분석의 원리

주성분분석의 원리

주성분이란 자료의 분산이 최대가 되는 방향을 나타내는 변수이다.

- 선형대수학을 건너뛰고 주성분(principal components)을 일단 직관적으로 이해해보자!
- 두 개의 변수 X_1과 X_2가 있다. 예를 들어 키와 몸무게를 생각해보자.
- 이 자료에서 분산이 가장 큰 "방향"을 찾으면 다음 그림에서 "직선 1"의 방향이다. 다른 방향(예컨대 "직선 2")에서 보면 자료의 분산이 작아짐을 눈대중으로 확인해보자.

주성분분석의 원리


Height
200
190
180
170
160
150
140
30
40
50
60
70
80
90
100
110
Weight
직선2
직선1

주성분분석의 원리

- 자료의 두 변수를 활용하여 “최대 분산을 설명하는” 변수(=벡터)를 새로 만들수 있다 (Why?).

$$PC_i = \omega_1 X_{1i} + \omega_2 X_{2i}$$

- 여기서 PC가 바로 주성분이고, ω는 요인적재량(factor loadings) 또는 요인부하량이다.
- 이 값들은 변수와 주성분 사이의 상관계수 r에 다름아니다(나중에 확인한다)

주성분분석의 원리

- 다시 말해, 주성분은 주어진 데이터에서 두 변수의 조합으로 생성된 변수(=벡터)인데, (아무 조합이라도 되는게 아니라) "분산을 극대화되는" 방향으로 만들어진다
- 주성분과 변수 X 및 변수 Y 사이의 요인적재량이 ω_1 과 ω_2 로 주어질 때, 그 제곱합 $(\omega_1^2 + \omega_2^2)$은 고윳값(eigenvalue)이 된다.
- 요인적재량 ω_1 과 ω_2 는 각각의 변수들이 주성분을 "얼마만큼 설명하는지"를 나타낸다.
- 그러므로 주성분의 고윳값이 크다는 것은 그 주성분을 잘 설명한 변수들이 제법 있었다는 것을 뜻한다.
- 뒤집어 말하자면, 고윳값이 큰 주성분이 있다는 사실은 하나의 새로운 변수(=주성분)만으로 다른 많은 기존의 변수들을 잘 설명하고 있다는 뜻이기도 하다.

주성분분석의 원리

계산된 주성분을 합성지수처럼 사용할 수 있다.

- 단순합 합성지수(composite index)를 만든다면 모든 요인부하량을 1로 통일시키는 것과 같다(Why?).
- 주성분을 계산할 때는 요인적재량을 (그대로 쓰는 대신) 고윳값으로 나누어 계산한다 (곧 연습한다). 이를 회귀기반 점수(regression-based scores)라고도 부른다.

주성분분석의 원리

지금까지의 과정을 Stata로 그대로 재현해보자.

- renpainters.dta에서 두 변수 composition와 expression 사이의 산점도를 확인하자.
- 두 변수를 가지고 주성분분석을 수행하고 하나의 주성분을 생성하자.
- 생성된 주성분과 두 변수 사이의 상관계수를 확인하고 주성분분석 결과표와 비교해보자.
- 요인적재량의 제곱합이 주성분의 고윳값과 일치하는지 여부도 확인하자.

두 개 이상의 주성분

두 개 이상의 주성분

물론 둘 이상의 주성분을 만들어낼 수 있다.

- 네 개의 변수 X_1, X_2, X_3, X_4가 있다고 하자.
- (아까와는 달리 그림을 그릴수 없는) 4차원의 공간에서 네 변수 사이에서 가장 큰 고윳값을 구할 수 있는 어떤 "방향"으로 첫번째 주성분(PC_1)을 만든다(직선1).
- 그리고 그 다음으로 큰 고윳값을 갖는 두번째 주성분(PC_2)을 만들 수 있다(직선2).

$$PC_1 = \omega_{11}X_1 + \omega_{12}X_2 + \omega_{13}X_3 + \omega_{14}X_4$$
$$PC_2 = \omega_{21}X_1 + \omega_{22}X_2 + \omega_{23}X_3 + \omega_{24}X_4$$

- 첫번째 주성분의 직선과 두번째 주성분의 직선 사이는 직교(orthogonal), 즉 90도 각도를 이루고 있다. 이것은 수학적인 함의를 가지고 있다.

두 개 이상의 주성분

- 아까와 마찬가지로 각각의 주성분과 네 개의 변수 X_1, X_2, X_3, X_4 사이 상관계수가 곧 요인적재량이다.
- 예컨대 첫번째 주성분에 대한 각 변수들의 요인적재량의 제곱합은 첫번째 주성분의 고윳값이 된다.

$$\omega_{11}^2 + \omega_{12}^2 + \omega_{13}^2 + \omega_{14}^2$$

- 큰 고윳값을 가진 주성분을 만들수 있었다는 것은 하나의 새로운 변수로 여러 변수들의 많은 변량(variation)을 설명할 수 있었다는 것을 뜻한다. 그만큼 차원은 축소된다.

두 개 이상의 주성분

- 주성분이 두 개 이상 생겨났을 때 제기되는 새로운 개념은 바로 공통분(communality)이다.
- 이는 하나의 변수가 두 개 이상의 주성분에 걸쳐 나누어 설명되는 요인적재량의 제곱합이다.
- 예컨대 변수 X_1의 첫번째 주성분에 대한 요인적재량과 두번째 요인적재량의 제곱합 $(= \omega_{11}^2 + \omega_{21}^2)$은 첫번째 변수의 공통분 값이다.
- 유일성(uniqueness)은 1에서 공통분(communality)을 뺀 값이다.

두 개 이상의 주성분

세 가지 기초 개념을 파악하는 것이 주성분분석의 출발점이다.

- (1) 요인적재량(factor loadings)은 주성분과 개별 변수들과의 상관계수이다.
- (2) 고윳값(eigenvalue)은 한 주성분이 다른 모든 변수들과 공유하는 분산이다.
- (3) 공통분(communality)은 특정 변수가 "주성분을 통해" 다른 모든 변수들과 공유하는 분산이다. 그러므로 다른 변수들과의 상관관계로 설명될 수 있다. 이는 결국 (앞서 설명한) 공통분산(common variance)의 추정치인 셈이다.
- "오차 분산(error variance)가 평균적으로 0이라는 가정 아래" 유일성(uniqueness)은 1에서 공통분산을 뺀 값이므로 이는 고유분산(unique variance)의 추정치인 셈이다.

주성분분석의 실제 응용

여러 개의 주성분이 나오면 어떻게 고를 것인가가 쟁점이 된다.

- 앞서 예제에서는 주성분이 하나와 두 개만 생성되었지만, 수많은 변수를 투입하면 주성분의 수도 그만큼 많아진다. 이를 반드시 모두 사용해야 하는 것은 아니다.
- 게다가 원변수의 수와 주성분의 수가 같아지면 주성분을 만든 의미가 없어진다 (Why?).
- 다만 요인을 추출하는 기준은 복잡하다. 사회통계학에서는 다음의 세 가지 기준이 주로 사용된다:

  (1) 카이저 규칙(Kaiser Rule), 즉 고윳값이 1보다 큰 주성분까지만 추출함.
  (2) 스크리 도표(scree plot)를 그려보고 팔꿈치 규칙(elbow rule)을 따름.
  (3) 기준 누적비율(cumulative proportion)에 도달하는 주성분까지만 추출함.

- 카이저 규칙은 근거가 부족하다는 설득력있는 비판에도 불구하고, 사회통계학 분야에서는 여전히 대세로 받아들여진다.

두 개 이상의 주성분

지금까지의 과정을 Stata로 그대로 재현해보자.

- 다시 renpainters.dta에서 네 변수 composition, drawing, colour, expression를 가지고 주성분분석을 수행하고 두 개의 주성분을 생성하자.
- 생성된 주성분들과 두 변수 사이의 상관계수를 확인하고 주성분분석 결과표와 비교해보자.
- 요인적재량의 제곱합이 주성분의 고윳값과 일치하는지 여부도 확인하자.
- 유일성을 계산하고 주성분분석 결과표와 일치하는지 여부도 확인하자.
- 누적비율(cumulative proportion)의 의미를 해석해보자.
- 두 개의 주성분이 갖는 실질적인 의미를 살펴보자.