

계량분석

상관분석

김현우, PhD¹

¹충북대학교 사회학과 조교수

October 25, 2021

진행 순서

- 1 공분산과 상관계수
- 2 유의성 검정 및 해석상 주의사항
- 3 상관분석의 실제 활용

공분산과 상관계수

공분산과 상관계수

둘 이상의 숫자형 변수 사이의 관계를 볼 때는 일차적으로 **상관분석** (correlation analysis)을 수행한다.

- 상관분석은 **상관계수(correlation coefficient)**를 구하는 과정이다.
- 상관계수를 이해하려면 먼저 **공분산(covariance)**을 이해할 필요가 있다.

다시 한 번 분산(variance)의 식을 돌이켜보자.

$$\begin{aligned}\text{Var}(\mathbf{x}) = \sigma^2 &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mu)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)\end{aligned}$$

공분산과 상관계수

변수가 하나(x_i) 주어져 있을 때 **편차(deviation)**를 구해 제공했는데 이것
이렇게 바꾸면 어떨까?

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

공분산과 분산의 아이디어는 거의 똑같다!

- 다만 x_i 하나의 편차를 제공하는 대신, x_i 와 y_i 의 편차를 서로 곱했을 뿐이다.
- 두 변수의 편차끼리 곱할 때, 모두 양수(+)이거나 음수(-)이면 양수(+)가 되고, 어느 한쪽이 양수(+)이고 다른 쪽이 음수(-)이면 음수(-)가 된다.

공분산과 상관계수

공분산을 Stata에서 직접 계산해보자([Stata 코드] 참고).

- eCampus에서 showmethemoney.csv를 다운받아 Stata로 불러오자. 이것은 단지 여덟 개의 관찰값만 가진 장난감 데이터(toy data)이다. 이 설문에 참여한 사람들은 월 소득(만원), 월 소득(백만원), 그리고 자신의 방 평수를 보고하였다.
- Stata에서 공분산을 계산하기 위해 사용해야 하는 명령어는 **correlate** 이고 옵션으로 반드시 **covariance**를 붙여야 한다.
- $Cov(income1, housesize)$, 즉 월 소득(만원)과 방 평수 사이의 공분산을 계산해보자.

공분산은 흥미로운 아이디어를 제시하고 있지만 명확한 단점이 있다.

- $\text{Cov}(X, Y)$ 는 **X 내부(within X)**의 분산과 **Y 내부(within Y)**의 분산이 다를 수 있다는 점을 고려하지 않는다. 제대로 표준화가 안되었다는 의미다.
- 그 결과 그 자체로는 해석이 안된다. 아니 대체 공분산이 241.12 이라고 나오면 그게 뭘 의미하나?

무슨 소리인지 잘 이해가 가지 않으니 직접 연습을 해보자!

- 아까 다운받은 showmethemoney.csv를 다시 열어보자.
- $\text{Cov}(\text{income1}, \text{housesize})$ 를 계산해 보고, 곧바로 $\text{Cov}(\text{income2}, \text{housesize})$ 를 계산해 보자.
- 결과가 같은가? 다른가? 정말로 income1과 income2이 다른 변수인가?

공분산과 상관계수

Karl Pearson은 이를 보완하는 천재적인 접근을 제시했다.

- 그는 두 변수 X와 Y의 각각의 표준편차(분산이 아니고!)를 분모로 각각 나누어줌으로서 X 내부의 분산과 Y 내부의 분산이 다를 수 있는 가능성을 제거하고 표준화를 이루었다.
- 뿐만 아니라, 일부러 분산이 아닌 표준편차로 나누어주었기 때문에 표준화된 값은 절묘하게 -1과 1사이로 두 변수가 얼마나 강한 상관관계를 가지고 있는지 보여준다.
- 이것이 이른바 **피어슨의 적률상관계수(Pearson's product-moment correlation coefficient)**이다. 줄여서 **상관계수(ρ)**다. ρ 는 rho라고 읽는다.
- 전에 언급한 바 있듯, σ 는 sigma라고 읽고 표준편차를 의미한다.

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

공분산과 상관계수

다시 showmethemoney.csv로 돌아가 상관계수를 계산해보자([Stata 코드] 참고).

- 먼저 $\text{Cov}(\text{income1}, \text{housesize})$ 를 계산해보자. 이때 income1의 표준편차와 housesize의 표준편차를 각각 기술통계량에서 계산할 필요가 있는데, Stata에서는 옵션으로 **means**를 덧붙이면 편리하다.
- 아래 식대로 엑셀에서 계산해보자. 괄호에 주의할 것!

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}$$

- 상관계수를 편리하게 구하기 위해 어떤 옵션도 붙이지 않고 **correlate** 명령어를 사용할 수 있다. 검산을 해보자.
- 이제 $\text{Corr}(\text{income1}, \text{housesize}) = \text{Corr}(\text{income2}, \text{housesize})$ 인지 확인하자.

유의성 검정 및 해석상 주의사항

유의성 검토 및 해석상 주의사항

상관계수의 해석은 매우 간단한테 혼동하지 않도록 주의해야 한다.

- 상관계수는 반드시 -1과 1 사이에 놓인다.
- 상관계수가 0보다 크면 두 변수는 **서로 같은 방향(정방향)**으로 움직인다. 즉 $\text{Cov}(x, y) > 0$ 이면 x가 커지면 y도 커진다. 반대로 0보다 작으면..
- 상관계수가 1에 가까울수록(그리고 -1에 가까울수록) 두 변수는 더욱 밀접한 관계를 갖게 된다.
- 해석은 이런 식이 편리하다: 0과 1 사이를 사분위수로 나누고 각각 리커트 4점 척도 (1사분위수=상관관계가 없다; 2사분위수=상관관계가 약하다; 3사분위수=상관관계가 어느 정도 있다; 4사분위수=상관관계가 강하다)로 의미를 부여하여 해석한다. 물론 0과 -1 사이에서도 마찬가지이다.

유의성 검토 및 해석상 주의사항

eCampus에서 homophily.dta를 다운받아 Stata에서 열자([Stata 코드] 참고).

- 이 자료는 장소와 시간대별로 성별-연령대별 유동인구수를 나타내 보이고 있다.
- 유유상종의 원리(the homophily principle)에 따라 거리를 활보할 때도 또래끼리 (20대는 20대끼리, 30대는 30대끼리 등) 함께 다닐 것이라고 가설을 세워 볼 수 있다.
- 예컨대 남자 10대가 가장 많이 다니는 시간-장소에 (마찬가지로) 많이 다니는 사람들은 누구인가?

모든 성별-연령대별 관계를 살펴보기 위해 상관계수행렬(correlation coefficient matrix)을 만든다.

- 두 개 이상의 변수를 입력하여 상관계수행렬을 쉽게 만들수 있다.
- 결과물을 하이라이트한 뒤, 우클릭-[Copy as Table]로 복사하여 엑셀에 붙여넣어 보자.

유의성 검토 및 해석상 주의사항

유의확률을 요약하기 위해 이제부터 별표(*)를 붙이기로 한다.

- p-value가 0.01보다 작으면 상관계수 옆에 별 2개(**), 0.05보다 작으면 별 1개(*), 0.1보다 작으면 대거 하나(†)을 붙여보자.
- 이것은 완전히 관습의 문제이고 심지어 사람마다 다르다. 어떤 사람은 아예 붙이지 않기도 한다.

Stata에서 상관계수의 유의성 검정을 하려면 `pwcorr` 명령어를 활용해야 한다([Stata 코드] 참고).

- **pwcorr**에서 pw는 pairwise를 뜻하며 변수를 두 개씩 짝지어(pairwise) 상관계수를 계산한다. 결측치가 없는 이상 **correlate**과 똑같은 결과를 보여준다(Why?).
- **sig** 옵션을 달면 p-value를 보여준다.
- **star(0.05)** 옵션을 달면 5%보다 작은 p-value에 대해 별표를 하나 붙여준다.

유의성 검토 및 해석상 주의사항

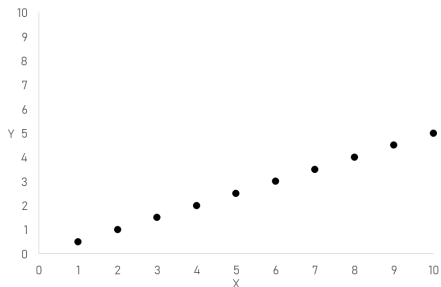
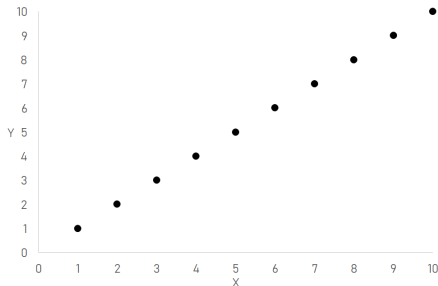
산포도와 상관계수는 늘 함께 생각하자!

- 산포도(scatterplot)를 통해 대략 상관계수가 어떻게 나올지 짐작할 수도 있다. 상관계수는 기본적으로 두 변수간 **선형적 관계의 강도(strength of the linear relationship)**를 나타내 보이기 때문이다.
- 의외로 기울기는 상관계수의 본질이 아니다! 오히려 관찰값 사이의 흩어짐(산포경향)이 상관계수의 본질이다(Why?).
- 참고로 기울기는 회귀계수의 본질을 전달한다. 혼동은 금물이지만 이에 관해서는 나중에 배우기로 한다.

유의성 검정 및 해석상 주의사항

- 왼쪽과 오른쪽 산포도 둘 다 상관계수(ρ)는 1이다.

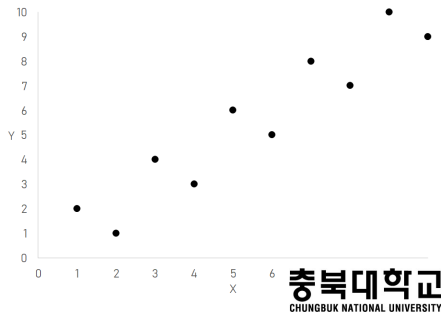
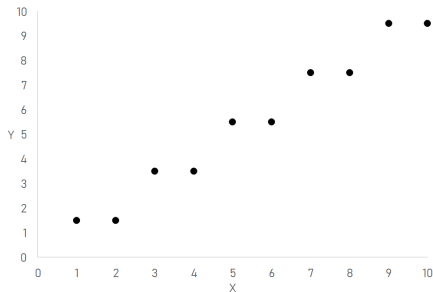
X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1	0.5	6	6	3
2	2	1	7	7	3.5
3	3	1.5	8	8	4
4	4	2	9	9	4.5
5	5	2.5	10	10	5



유의성 검정 및 해석상 주의사항

- 왼쪽 산포도에서 Y는 X로부터 ± 0.5 씩 더해 계산되었고, 오른쪽 산포도에서 Y는 X로부터 ± 1 씩 더해 계산되었다.
- 왼쪽 상관계수($\rho = .98$)가 오른쪽 상관계수($\rho = .94$)보다 크다.

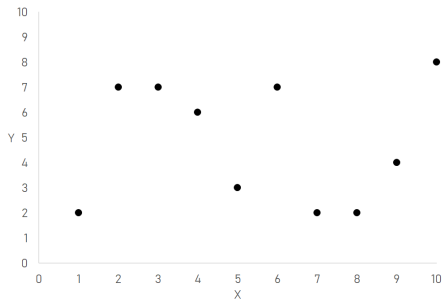
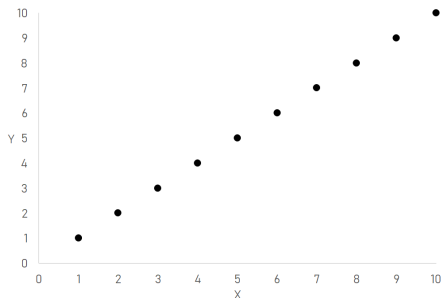
X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1.5	2	6	5.5	5
2	1.5	1	7	7.5	8
3	3.5	4	8	7.5	7
4	3.5	3	9	9.5	10
5	5.5	6	10	9.5	9



유의성 검정 및 해석상 주의사항

- 왼쪽 산포도의 상관계수(ρ)이지만 (규칙없이 퍼진) 오른쪽 산포도의 ρ 는 0이다.

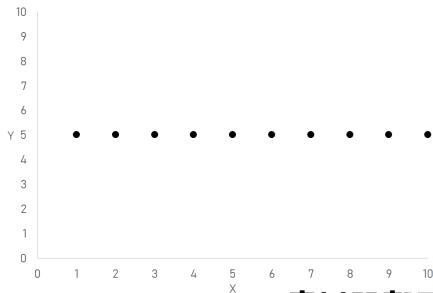
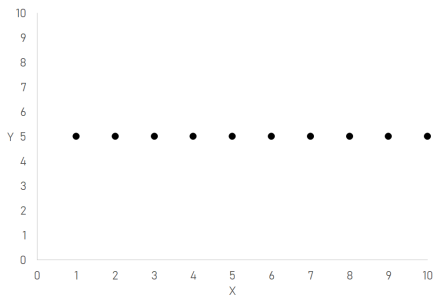
X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	1	2	6	6	7
2	2	7	7	7	2
3	3	7	8	8	2
4	4	6	9	9	4
5	5	3	10	10	8



유의성 검정 및 해석상 주의사항

- 왼쪽 산포도에서 Y는 .0001씩이라도 커져서 $\rho = 1$ 이지만, 오른쪽 산포도는 전혀 변동하지 않아 수평선이고 ρ 는 아예 계산되지 않는다(Why?).

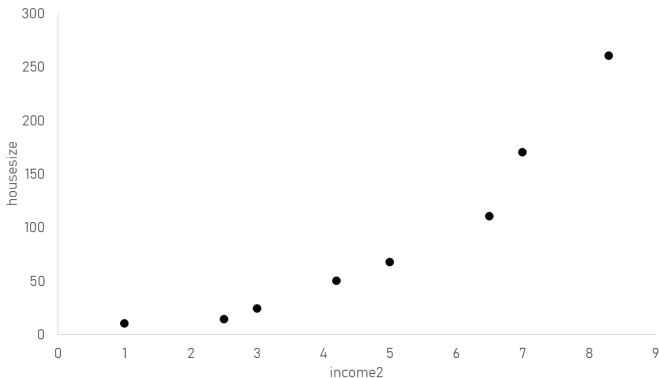
X(공통)	Y(왼쪽)	Y(오른쪽)	X(공통)	Y(왼쪽)	Y(오른쪽)
1	5	5	6	5.0005	5
2	5.0001	5	7	5.0006	5
3	5.0002	5	8	5.0007	5
4	5.0003	5	9	5.0008	5
5	5.0004	5	10	5.0009	5



유의성 검정 및 해석상 주의사항

아까 showmethemoney.csv에서 사실 월 소득과 방 평수 사이의 관계는 **비선형적(non-linear)**이다.

- 이것은 선형적(linear)이지 않고 보다 구체적으로는 **곡선형(curvilinear)**이다.



유의성 검토 및 해석상 주의사항

상관계수를 보고할 때는 반드시 함께 산포도를 그려보자([Stata 코드] 참고).

- X와 Y의 관계는 경우에 따라서 U자형, 역U자형, W자형 등등 다양할 수도 있다.
- 상관분석은 기본적으로 X와 Y의 관계가 **선형적일 것으로 가정**하기 때문에 만일 데이터가 그렇지 않으면 문제를 일으킨다.
- 따라서 무턱대고 상관계수를 보고하기 보다는 반드시 먼저 산포도를 그려보고 충분히 선형적 관계인지 눈으로 확인해 둘 필요가 있다.
- Stata에서도 상관계수를 보고하기 앞서 산포도를 그려보자. 산포도에서 찾은 극단치(outliers)를 제거하고 다시 상관계수도 계산해보자.

상관분석의 실제 활용

상관분석의 실제 활용

상당히 많은 논문들이 기술통계의 일환으로 상관계수행렬을 제시한다.

- 상관계수행렬은 단순히 한 변수와 다른 변수 사이의 상관관계를 보여주는 것을 넘어 이른바 **다중공선성(multicollinearity)**이 존재하는가를 살펴볼 때도 (잠재적으로) 유용하다. 이에 관해서는 나중에 다루기로 한다.
- 상관계수행렬은 논문 한 페이지를 통째로 잡아먹기 때문에 근래에는 보고하지 않는 경우도 많다.
- 이 행렬은 이른바 **메타분석(meta analysis)**에서 종종 유용한 정보를 제공하는 경우가 있으므로 사정이 허용된다면 온라인 부록(online appendix)으로라도 보고하는 편이 바람직하다.

논문의 기술통계 파트에서 상관계수행렬이 실제로 어떻게 활용되는지 살펴보자

- 한내창 (2020)의 <표2>를 꼼꼼히 살펴보자. 이 표는 혼전성수용도에서 교육수준에 이르기까지 10개의 변수들 사이에 상관관계수가 어떠한가를 보여준다.
- 혼전성수용도와 가장 큰 상관계수를 보이는 변수는 무엇인가?
- 혼외성수용도와 가장 큰 상관계수를 보이는 변수는 무엇인가?
- 혼전 및 혼외성수용도 사이에는 통계적으로 유의한 상관관계가 있나?
- 가장 큰 상관계수를 보이는 두 변수는 무엇인가?

충북대학교
CHUNGBUK NATIONAL UNIVERSITY