사회통계
상관분석

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 상관분석 톺아보기
2. 상관계수 유의성 검정
3. 상관분석의 실제 활용

상관분석 톺아보기

상관분석 톺아보기

두 양적변수 사이의 관계를 살펴볼 때는 상관분석을 수행한다.

- 상관분석(correlation analysis)은 상관계수(correlation coefficient)를 구하는 기법이다.
- 상관계수를 이해하려면 먼저 분산(variance)과 공분산(covariance)을 돌이켜 볼 필요가 있다.
- 먼저 분산의 식을 돌이켜보자.

$$Var(X) = \sigma_X^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \mu)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)(X_i - \mu)$$

- 변수가 하나 주어져 있을 때 편차(deviation)를 구해 제곱하여 분산을 구했는데, 아래처럼 살짝 바꾸면 곧바로 공분산이 된다.

$$Cov(X, Y) = \sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

상관분석 톺아보기

- 공분산과 분산의 아이디어는 거의 똑같다! 다만 X_i 하나의 편차를 제곱하는 대신, X_i와 Y_i의 편차를 서로 곱했을 뿐이다.
- 두 변수의 편차끼리 곱할 때, 모두 양수(+)이거나 음수(-)이면 공분산은 양수(+)가 되고, 어느 한쪽이 양수(+)이고 다른 쪽이 음수(-)이면 공분산은 음수(-)가 된다.
- 엑셀에서 COVARIANCE.P(·) 함수로 모집단의 공분산을 구할 수 있다. 표본의 공분산은 COVARIANCE.S(·) 함수로 구할 수 있다.
- 공분산은 흥미로운 아이디어를 제시하고 있지만 명확한 단점이 있었다.
- $Cov(X, Y)$는 X 내부(within X)의 분산과 Y 내부(within Y)의 분산이 다를 수 있다는 점을 고려하지 않았으므로 그 자체로는 해석이 어려웠다.

상관분석 톺아보기

Karl Pearson은 공분산의 단점을 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 ρ다. ρ는 rho라고 읽는다.

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

상관분석 톺아보기

상관계수의 해석은 매우 간단하지만 혼동하지 않도록 주의해야 한다.

- 상관계수는 반드시 -1과 1사이에 놓인다. $\rho = 0$는 무상관(no correlation), $\rho = 1$은 완전상관(perfect correlation)이라고 한다.
- 상관계수가 0보다 크면 두 변수는 서로 같은 방향(정방향)으로 움직인다. 즉 "X가 증가하면 Y도 증가한다."
- 상관계수가 0보다 작으면 두 변수는 서로 다른 방향(역방향)으로 움직인다. 즉 "X가 증가하면 Y는 감소한다."
- 왜 이렇게 해석되는지 공분산의 분자 $\sum(x - \mu_X)(y - \mu_Y)$를 잘 들여다보자.
- 상관계수가 1에 가까울수록 (그리고 -1에 가까울수록) 두 변수는 더욱 밀접한 상관관계를 갖게 된다.

상관분석 톺아보기

- 해석할 때는 0과 1 사이를 사분위수로 나누고 각각 리커트 4점 척도로 의미를 부여한다. 물론 0과 -1 사이에서도 마찬가지이다.

상관계수	상관관계의 해석
[-1, -0.75]	매우 강한 역방향
[-0.75, -0.5]	다소 강한 역방향
[-0.5, -0.25]	다소 약한 역방향
[-0.25, 0]	매우 약한 역방향
[0, 0.25]	매우 약한 정방향
[0.25, 0.5]	다소 약한 정방향
[0.5, 0.75]	다소 강한 정방향
[0.75, 1]	매우 강한 정방향

상관분석 톺아보기

예제 1. fullauto.csv를 사용하여 차체회전반경(turn)과 전장(length) 간의 상관계수를 구하고 이를 해석하시오. 상관관계를 나타내는 그래프를 함께 제시하시오.

상관분석 톺아보기

- 먼저 두 변수의 의미를 코드북(codebook)에서 파악하자.
- 이제 두 변수의 자료유형을 확인해야 한다. 빈도분포표 혹은 히스토그램을 그려서 이를 확인하자.
- 두 변수는 양적변수이므로 엑셀 함수로는 CORREL(·) 혹은 PEARSON(·)을 사용하여 상관계수를 구한다.
- 상관계수에 따르면 두 변수 사이에는 상관관계가 있는가? (있다면) 어느 방향으로 얼마나 강한 상관관계가 있는가?
- 물론 앞서 제시한 식을 직접 이용할 수도 있다. 두 변수 사이의 공분산을 구한 다음, 각 변수의 표준편차의 곱으로 나누어주면 마찬가지로 상관계수를 구할 수 있다.

상관계수 유의성 검정

상관계수 유의성 검정

상관계수에 대해서도 유의성 검정을 할 수 있다.

- 상관계수에 대해서는 대체로 양측검정을 수행하므로 가설 구조는 다음과 같다.

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

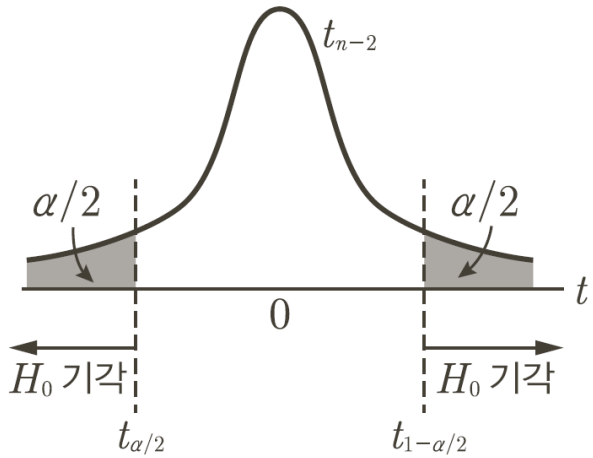
- 귀무가설이 참이라는 가정 아래 무한히 계속 표본을 뽑아 그것들의 상관계수 r을 구해보면 이것은 t 분포를 따른다.
- 이 t 분포의 꼴은 $n-2$의 자유도로 결정된다(Why?).
- 필요에 따라 $\rho > 0$ 이나 $\rho < 0$ 같은 단측검정을 수행할 수도 있지만 좀처럼 쓰이지 않는다.

상관계수 유의성 검정

- 검정통계량 t 값을 구하고 유의확률(p-value)을 계산한다.

$$t = \frac{r-\rho}{\sigma_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = r\sqrt{\frac{n-2}{1-r^2}}$$

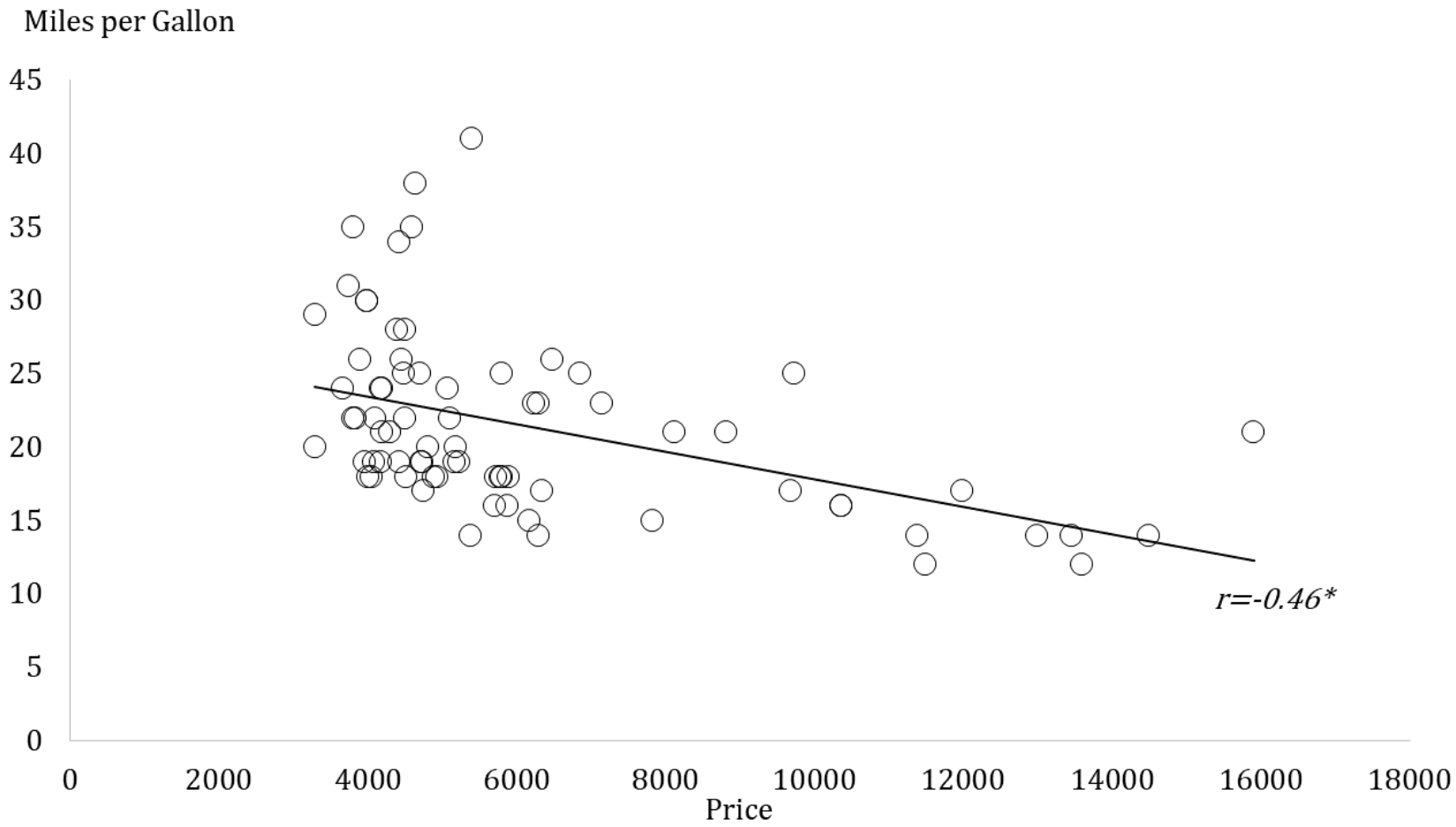
- 이때 σ_r은 표본상관계수의 표준오차(standard error)이다.
- 즉 (1) 표본상관계수 r가 커지고 (2) 표본 크기 n이 커질수록 t 값이 커져 귀무가설을 기각하기 쉬워진다.



상관계수 유의성 검정

예제 2. fullauto.csv를 다시 사용하여 차량 가격(price)과 가성비(mpg) 간의 상관계수를 구하고 이를 해석하시오. 상관관계를 나타내는 그래프를 함께 제시하시오. 또한 99.9% 신뢰수준에서 가설검정을 수행하시오.

상관계수 유의성 검정


Miles per Gallon
45
40
35
30
25
20
15
10
5
0
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
Price
r=-0.46*

상관계수 유의성 검정

- 양측검정을 수행하며, 주어진 가설 구조는 다음과 같다:

$$H_0 : \rho = 0$$
$$H_a : \rho \neq 0$$

- 엑셀에서 CORREL$(\cdot)$ 함수로 두 변수 사이의 상관계수를 구한다.

$$r = -0.459$$

- 검정통계량 t 값은 다음과 같다.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -4.389$$

상관계수 유의성 검정

- t 분포의 양쪽 꼬다리의 "곡선 밑 면적"을 구하고 싶다(Why?).
- 그러므로 엑셀에서 T.DIST(-4.389, 72, TRUE)와 1-T.DIST(4.389, 72, TRUE)를 입력해 나온 값들을 더해 유의확률(p-value)을 구한다.
- $p < 0.001$ 이므로 99.9% 신뢰수준에서 통계적으로 유의하게 귀무가설을 기각한다.
- 정말로 외제차 가성비와 가격 사이에는 다소 약한 역(-)의 상관관계가 존재하는 것 같다!

상관계수 유의성 검정

유의확률에 관한 정보를 요약하기 위해 이제부터 *을 붙이기로 한다.

- 유의확률이 0.001보다 작으면 상관계수 옆에 별 3개(***), 0.01보다 작으면 별 2개(**), 0.05보다 작으면 별 1개(*), 0.1보다 작으면 대거(dagger) 하나(†)를 붙일 수 있다.
- 이런 표식은 결국 "통계적으로 유의하게 귀무가설을 기각할 수 있음"을 의미한다.
- 이것은 완전히 관습의 문제이고 심지어 사람마다 다르다. 어떤 사람은 아예 붙이지 않기도 한다.
- 아까 나의 그림에서는 그냥 유의하다는 사실을 알리기 위해 별 하나로 퉁쳤다.

상관계수 유의성 검정

소표본만 벗어나도 상관계수의 유의성 검정에 사실 큰 의미가 없다.

- 위 식을 곰곰히 들여다보면 눈치챌 수 있는 부분인데, 예컨대 $n = 50$ 정도로 작은 샘플에서 $r = 0.3$ 정도의 값만 나와주어도 이미 95% 신뢰수준에서 통계적으로 유의하다(Why?).
- 오늘날 경험적 사회과학 연구에서 $n = 50$ 짜리 연구는 없다. 게다가 $r = 0.3$는 거의 없는 수준의 상관관계에 불과하다.
- 그러다보니 상관분석의 경우에는 구태여 별을 붙이지 않는 경우도 있다.

상관분석의 실제 활용

상관분석의 실제 활용

상관계수를 보고할 때는 반드시 함께 산점도를 그려야 한다.

- 상관계수는 기본적으로 두 변수간 선형적 관계의 강도(strength of the linear relationship)를 나타내 보인다.
- 다시 말해, 두 변수 사이에 선형적이지 않은 관계, 즉 비선형적 관계(nonlinear relationship)가 있는 경우에는 상관계수가 오해를 불러온다.
- 애시당초 두 변수 사이가 U자형, 역U자형, W자형 등등이 아니라는 보장이 어디에 있을까?

상관분석의 실제 활용

예제 3. fullauto.csv를 사용하여 후방좌석길이(rseat)과 배기량(displ) 사이의 상관계수 r을 확인하고 산점도(scatterplot)를 그리시오.

상관분석의 실제 활용

- 극단치(outliers)가 있는 경우 상관계수는 여기에 민감하게 영향받음을 알 수 있다.
- 그러므로 (극단치의 존재를 식별해내기 위해서라도) 반드시 산점도를 그려보아야 한다.
- 필요하다면 극단치를 제거한 뒤에 다시 상관계수를 계산하는 것이 바람직할 수도 있다(물론 제거 여부를 꼭 밝혀두어야 한다).

상관분석의 실제 활용

모든 변수들 사이의 관계를 한 번에 살펴보기 위해 상관계수행렬을 만든다.

- 원칙적으로 숫자형 변수가 두 개 사이의 관계를 볼 때 상관분석을 수행한다. 하지만 여러 개 있어도 큰 문제가 없다.
- 만약 10개의 변수가 있으면 쌍대비교(pairwise comparison)를 45번($= {}_{10}C_2$)하면 그만이다(Why?).
- 그렇지만 45번을 매번 따로따로 보고하면 좀 보기 흉할 것 같다.
- 그러므로 여기서는 차라리 변수들을 쭉 나열한 뒤, 상관계수를 요약해서 보고하는 상관계수행렬(correlation coefficient matrix)을 한 번에 만드는 쪽이 낫다.

상관분석의 실제 활용

예제 4. fullauto.csv를 사용하여 모든 숫자형 변수들 price, mpg, hdroom, rseat, trunk, weight, length, turn, displ, gratio의 상관계수행렬을 만드시오.

상관분석의 실제 활용

- 변수는 모두 10개이므로 45개의 상관계수를 보고하게 된다.
- 엑셀에서 [데이터]-[데이터 분석]을 통해 “상관 분석”을 선택한다. “첫째 행 이름표 사용”을 고려하여 적절하게 자료를 하이라이트한다.
- 나온 결과에서 셀의 크기와 소숫점을 적절히 컨트롤한다.
- 많은 연구 보고서와 논문에서는 기술통계(descriptive statistics)의 일환으로 상관계수행렬을 제시하는 편이다.
- 다만 상관계수행렬은 한 페이지를 통째로 잡아먹기 때문에 근래에는 보고하지 않는 경우도 많아졌다.
- 여기에 더해 상관계수 옆에 별을 잔뜩 붙이다보면 페이지 공간을 쓸떼없이 더 차지하고 유의성 검정이 사실 별 의미도 없다.

상관분석의 실제 활용

학술논문을 통해 상관계수행렬이 실제로 어떻게 활용되는지 살펴보자.

- 한내창 (2020)의 〈표1〉을 꼼꼼히 살펴보자.
- 비본질적 종교성의 측정도구로 몇 가지 측정문항을 사용하였는가? 그것들은 각각 무엇인가?
- 다음으로 〈표2〉 표는 혼전성수용도에서 교육수준에 이르기까지 10개의 변수들 사이에 상관계수가 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?

한내창. 2010. “종교와 성태도 간 관계.” 한국사회학 44(5): 114-138.

상관분석의 실제 활용

〈표 1〉 분석 모델에 포함된 변인들의 간략한 기술적 내용

구분	변수명	문항수	값	평균/SD	α
종속 변수	1. 혼전성수용도	1	1- 4	2.40/ 1.05	-
	2. 혼외성수용도	1	1- 4	1.39/ 0.73	-
독립 변수	3. 신봉종교	1	-	-	-
	4. 종교서비스참여	1	월평균빈도	1.69/ 2.61	-
	5. 기도빈도	1	월평균빈도	10.28/19.51	-
	6. 기타활동참여	1	월평균빈도	0.88/ 2.03	-
	7. 주관적종교성	1	1- 7	3.94/ 1.77	-
	8. 본질적종교성	1	1- 5	2.27/ 1.33	-
	9. 비본질적종교성	4	4-20	15.41/ 3.02	0.81
통제 변수	10. 성	1	-	-	-
	11. 연령(18-91)	1	만 나이	45.73/34.82	-
	12. 교육수준	1	교육년수	12.43/ 4.26	-

1. 남녀가 결혼 전에 성관계를 갖는 것이 옳다고 생각하십니까?
2. 결혼한 사람이 배우자가 아닌 사람과 성관계를 갖는 것이 옳습니까?
3. 귀하는 어떤 종교를 가지고 계십니까?
4. 귀하는 얼마나 자주 불공 또는 예배드리러 가십니까?
5. 귀하는 얼마나 자주 기도하십니까?
6. 귀하는 현재 종교의식(예배나 법회 등)에 참석하는 것 외에 교회, 성당, 절 등에서 하는 모임이나 활동에 얼마나 자주 참여하십니까?
7. 귀하는 자신이 얼마나 종교적이라고 생각하십니까?
8. 나에겐 오직 신이 존재하기 때문에 삶이 의미가 있다.
9. 종교생활을 하는 것은 내적 평화와 행복을 얻는데 도움이 된다
   종교생활을 하는 것은 친구를 사귀는데 도움이 된다
   종교생활을 하는 것은 어렵거나 슬플 때 위안을 얻는데 도움이 된다
   종교생활을 하는 것은 나와 잘 맞는 사람을 만나는데 도움이 된다

상관분석의 실제 활용

〈표 2〉 주요 변인들 간 피어슨 단순 상관계수

	1	2	3	4	5
1. 혼전성수용도	-				
2. 혼외성수용도	.34***	-			
3. 법회·예배참석	-.22***	-.09***	-		
4. 기도빈도	-.18***	-.10***	.63***	-	
5. 기타활동 참석	-.15***	-.07**	.68***	.53***	-
6. 자기평가종교성	-.17***	-.06**	.58***	.52***	.44***
7. 본질적종교성	-.20***	-.09***	.58***	.53***	.44***
8. 비본질적종교성	-.02	-.04	.35***	.33***	.30***
9. 연령	-.41***	-.09***	.14***	.15***	.07**
10. 교육수준	.29***	.09***	-.01	-.02	.04

※ 유의수준: ***p 〈0.001 **p 〈0.01 *p 〈0.05

〈표 2: 계속〉 주요 변인들 간 피어슨 단순 상관계수

	6	7	8	9
6. 자기평가종교성	-			
7. 본질적종교성	.56***	-		
8. 비본질적종교성	.40***	.34***	-	
9. 연령	.14***	.16***	.00	-
10. 교육수준	-.05*	-.08**	.09***	-.62***

※ 유의수준: ***p 〈0.001 **p 〈0.01 *p 〈0.05