13.7 상관관계 만들어 내기

학습지

문제

STEP III 2010, 13번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/10-s3-q13

힌트

먼저 혼자 힘으로 문제를 풀어 보자. 막히면 힌트를 **하나씩** 읽고, 힌트를 하나 읽을 때마다 다시 문제로 돌아가자. **잠깐, 직접 해 보기**라고 표시된 힌트가 핵심 단계이다. 각 힌트 뒤에는 풀이를 적을 공간이 있다.

? **생각해 보기** Z_1과 Z_2는 독립이다. 둘의 공분산은 얼마이고, 따라서 상관계수는 얼마인가?

1단계 두 변수

다루는 내용: Z_1과 Z_2의 상관계수, 그리고 변수 Y_2

? **생각해 보기** 독립인 변수들의 공분산(covariance)은 0이다. Y_2에 대해서는 공분산의 선형성(linearity)과 Z_1, Z_2의 독립성(independence)을 이용하여 $\mathrm{Var}(Y_2)$와 $\mathrm{Cov}(Y_1, Y_2)$를 구한다.

2단계 세 번째 변수

다루는 내용: 변수 Y_3

잠깐, 직접 해 보기 $\mathrm{Var}(Y_3) = 1, \mathrm{Cov}(Y_1, Y_3) = \rho_{13}, \mathrm{Cov}(Y_2, Y_3) = \rho_{23}$을 a, b, c에 대한 방정식으로 쓴다. 이 방정식들을 a, b, c의 순서로 푼다.

3단계 임의의 평균과 분산

다루는 내용: 마지막 부분

> **생각해 보기** 변수를 평행이동하고 상수배하면 그 평균, 분산, 상관계수(correlation)는 각각 어떻게 되는가? $X_i = \mu_i + \sigma_i Y_i$로 놓아 본다.

숨은 수학

숄레스키 분해

행렬 형태로 쓰면, 이 구성은 **하삼각**(*lower triangular*) 행렬

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & \sqrt{1-\rho_{12}^2} & 0 \\ a & b & c \end{pmatrix},$$

에 대해 $\mathbf{Y} = L\mathbf{Z}$라는 것이고, $\mathbf{Y}$의 상관행렬(correlation matrix)은 $R = LL^{\mathsf{T}}$이다. 대칭행렬을 LL^{T} 꼴로 나타내는 것이 **숄레스키 분해**이다. 이 분해는 프랑스의 군사 기술자 앙드레루이 숄레스키(André-Louis Cholesky)가 1910년 무렵 측지 측량의 방정식을 풀기 위해 찾아냈다(그가 제1차 세계대전에서 사망한 뒤에 발표되었다). 이 분해는 R가 **양의 정부호**(*positive definite*)이면 언제나 가능하며, 그 계산은 우리가 손으로 한 것과 정확히 같다. 각 행을 차례로 풀면 된다. 이것은 금융, 공학, 기후 모델링에서 상관된 변수를 모의실험하는 표준적인 방법이다.

어떤 상관계수가 가능한가?

c의 제곱근 안에 있는 수는 음이 아니어야 한다. 전개하면 그 조건은

$$1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23} \geq 0,$$

이고, 이는 상관행렬의 행렬식(determinant)이 음이 아니라는 조건이다. 기하학적으로 보면, 두 개씩 이루는 각이 θ_{ij}(여기서 $\cos\theta_{ij} = \rho_{ij}$)인 세 단위벡터가 실제로 공간 안에 함께

놓일 수 있어야 한다는 뜻이다. $\rho_{12} = \rho_{13} = \rho_{23} = -0.6$이면 각은 모두 약 127°이고, 서로 127°를 이루는 세 벡터는 각의 합이 360°보다 커야 하므로 불가능하다. 일반적으로 어떤 행렬이 상관행렬이 될 수 있으려면 양의 준정부호(positive semi-definite), 즉 모든 고윳값(eigenvalue)이 음이 아니어야 한다(Section 7.2). 불완전한 자료로부터 상관계수를 하나하나 따로 추정하면 이 조건이 깨질 수 있으며, 그러면 통계학자는 가장 가까운 유효한 상관행렬을 찾아야 한다. 이 문제는 2002년에 Nicholas Higham이 해결했다.

골턴, 피어슨, 그리고 상관

프랜시스 골턴(Francis Galton)은 1888년 부모와 자녀의 키가 어떻게 관련되는지를 연구하면서 상관이라는 개념을 도입했고, 칼 피어슨(Karl Pearson)은 1896년 이 문제에서 사용한 적률 공식(product moment formula)을 제시했다. 골턴은 평균으로의 회귀(regression to the mean)도 발견했다(Section 13.1). 피어슨은 1911년 유니버시티 칼리지 런던(University College London)에 세계 최초의 대학 통계학과를 세웠다. "상관관계는 인과관계(causation)를 뜻하지 않는다"라는 경고는 상관이라는 개념 자체만큼이나 오래되었다.

더 도전하기

1 **불가능한 상관계수.** 세 확률변수의 두 개씩의 상관계수가 모두 ρ이려면 $\rho \geq -\frac{1}{2}$이어야 함을 보여라. n개의 변수에 대해서는 이에 대응하는 명제가 무엇인가?

2 **행렬식 확인하기.** R가 3×3 상관행렬일 때, $1 - \rho_{13}^2 - b^2 \geq 0$이 $\det R \geq 0$과 동치임을 보여라.

3 **포트폴리오 위험.** 두 주식의 수익률의 표준편차(standard deviation)가 $\sigma_1 = 0.2$, $\sigma_2 = 0.3$이고 상관계수가 ρ이다. 두 주식에 절반씩 투자한 포트폴리오의 표준편차를 $\rho = -1, 0, 1$일 때 각각 구하여라. 분산투자(diversification)는 어떤 효과를 내는가?

4 **재료 되찾기.** 문제에서와 같이 Y_1, Y_2가 주어졌을 때, Z_2를 Y_1과 Y_2로 나타내어라. (이것이 그람–슈미트 단계이다.)

더 읽을거리

- **길버트 스트랭(Gilbert Strang),** ***Introduction to Linear Algebra*** (Wellesley-Cambridge). 양의 정부호 행렬과 숄레스키 분해를 다룬다.
- **Paul Glasserman,** ***Monte Carlo Methods in Financial Engineering*** (Springer). 상관된 변수를 실제로 모의실험하는 방법을 다룬다. 수준이 높지만 앞부분의 장들은 읽을 만하다.

► **Theodore Porter, *Karl Pearson: The Scientific Life in a Statistical Age*** (Princeton University Press).

안내에 따른 전체 풀이는 제 13.7절에 있다:
https://beyondstep.keithmcnulty.org/ko/ch13/correlation.html

매주 STEP 문제 하나를 끝까지 풀어 본다. The STEP Up은 이메일로 무료로 받아 볼 수 있다: keithmcnulty.substack.com