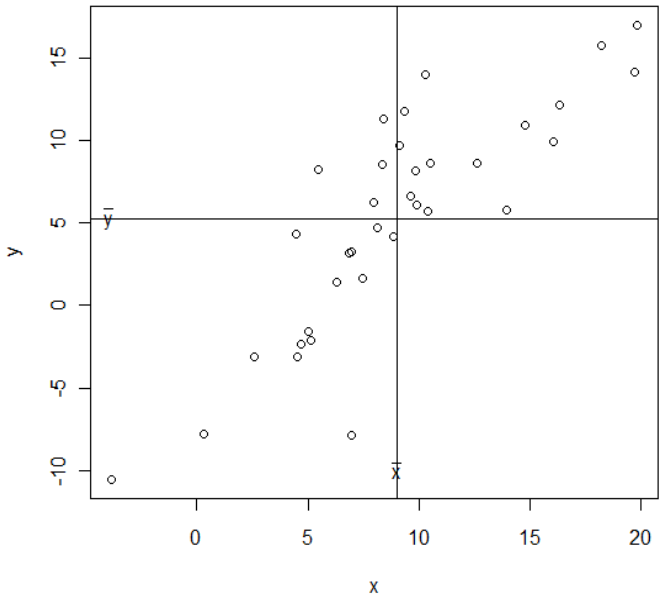
SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

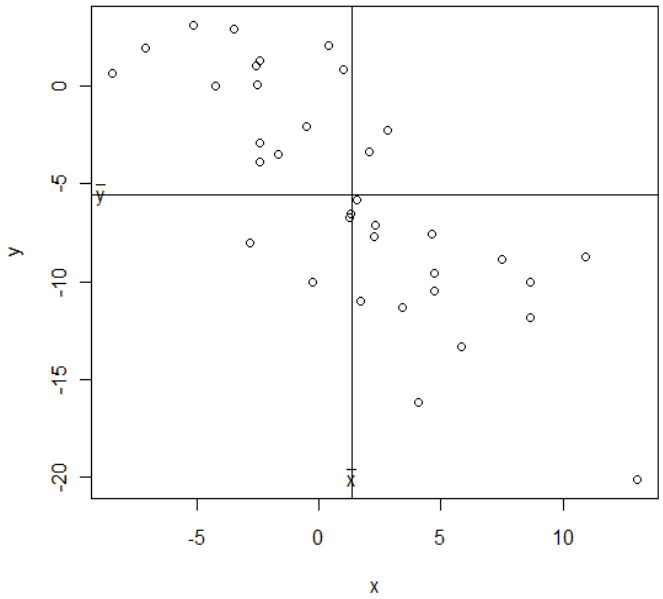
■ 공분산과 상관계수

- 산점도: 두 수치변수 간에 관계가 있는지를 시각적으로 확인
- 두 수치변수 간에 **직선관계**가 어느 정도인지를 나타내는 통계값
- 자료표시: $(x_1, y_1), (x_2, y_2), \ \ldots\ , (x_n, y_n)$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

◉ 양(좌)과 음(우)의 관계를 가지는 산점도





- 고려사항
  - 위치에 따라 직선관계에는 변화가 없음 ⇒ $(\overline{x}, \overline{y})$를 중심으로
  - 좌 그림: $(\overline{x}, \overline{y})$를 중심으로 1과 3사분면에 자료가 많고 길게 분포
    ⇒ 양수로 표시
  - 우 그림: $(\overline{x}, \overline{y})$를 중심으로 2와 4사분면에 자료가 많고 길게 분포
    ⇒ 음수로 표시
  - $(\overline{x}, \overline{y})$에서 멀어질수록 직선관계가 명확해짐

⇨ $(x_i - \overline{x})(y_i - \overline{y})$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 표본공분산(sample covariance)

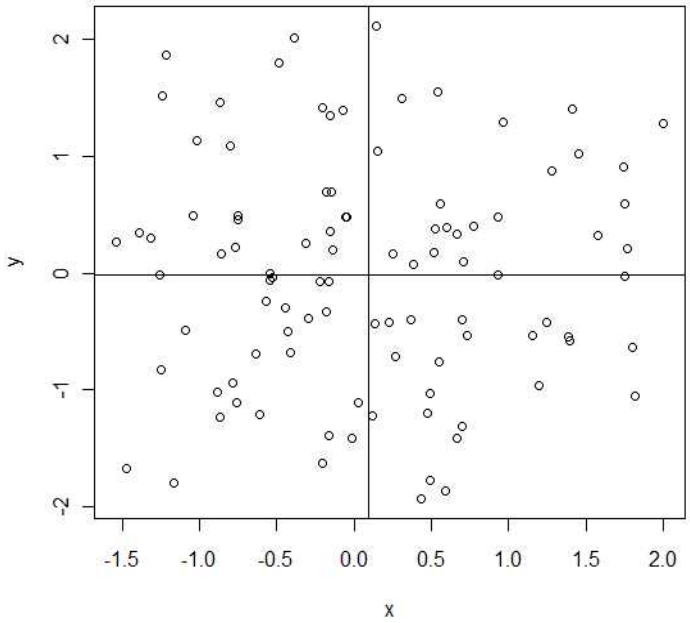
$$c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

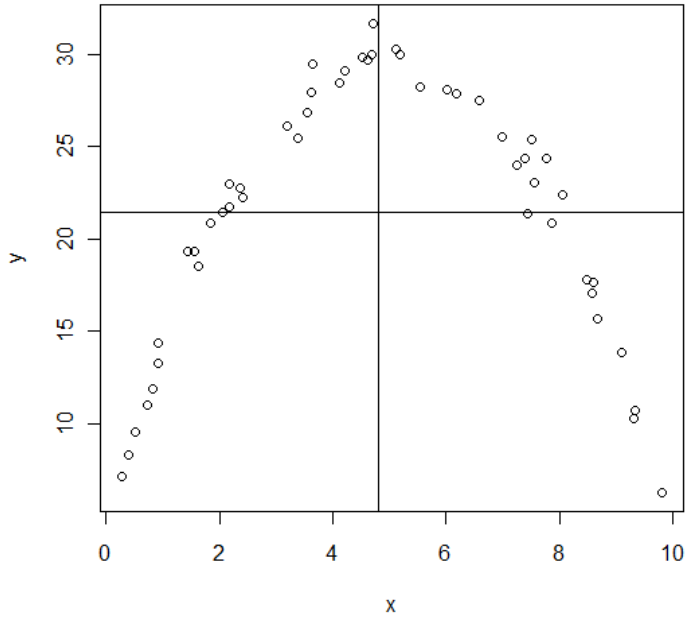
- 좌 그림: 양의 기울기인 선분에 자료가 모여 있음 $\Rightarrow$ $c > 0$
- 우 그림: 음의 기울기인 선분에 자료가 모여 있음 $\Rightarrow$ $c < 0$
- y_i를 x_i로 바꾸면

$$c = \frac{1}{n-1}\sum(x_i - \overline{x})(x_i - \overline{x}) = \frac{1}{n-1}\sum(x_i - \overline{x})^2$$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

◉ 직선관계가 없는 산점도 ($c \approx 0$)





- 표본공분산의 간편식

$$
\begin{aligned}
c &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - n\overline{x}\,\overline{y}\right\} \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}x_i\,y_i - \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\right\}
\end{aligned}
$$

◉ 올림픽 육상 100미터 우승기록

○ 남자(1900~2016년 자료)

번호	x(연도)	y(기록)	x^2	y^2	xy
1	1900	11	3610000	121	20900
2	1904	11	3625216	121	20944
⋮	⋮	⋮	⋮	⋮	⋮
27	2016	9.81	4064256	96.236	19776.96
합	52940	276.76	103835088	2841.27	542291.2

○ 연도와 남자 우승기록의 표본공분산

$$c = \frac{1}{27-1}\left(542291.2 - \frac{1}{27}(52940)(276.71)\right) = \frac{-363.45}{26} = -13.98$$

○ 여자

번호	x(연도)	y(기록)	x^2	y^2	xy
1	1928	12.2	3717184	148.84	23521.6
2	1932	11.9	3732624	141.61	22990.8
⋮	⋮	⋮	⋮	⋮	⋮
21	2016	10.71	4064256	114.794	21591.36
합	41472	234.38	81915488	2619.86	462655.6

○ 연도와 여자 우승기록의 표본공분산

$$c = \frac{1}{21-1}\left(462655.6 - \frac{1}{21}(41472)(234.38)\right) = \frac{-211.457}{20} = -10.57$$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 표본상관계수(coefficient of correlation)

- 표본공분산의 문제점
  - ○ 측정 단위에 영향을 받기 때문에 그 값 자체로 선형관계의 정도를 알 수는 없음
  - ○ 예】우승기록을 초 ⇒ 분 단위로 표시
    ⇨ 남자의 표본공분산: –13.98/60=-0.233

- 피어슨의 표본상관계수
  - 표준화된 자료의 표본공분산

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)$$

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

- 표본상관계수의 간편식

$$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_i y_i-n\bar{x}\,\bar{y}$$

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2$$

$$S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$$

⇨

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

◦ Cauchy-Schwartz 부등식: $\left(\sum a_i b_i\right)^2 \le \sum a_i^2 \sum b_i^2$

$$\Rightarrow |r| \le 1 \;\rightleftarrows\; -1 \le r \le 1$$

- 표본상관계수의 성질
  - 기울기를 가지는 직선에 조밀하게 모일수록 $|r|$는 1에 근접
    - 모든 관측값들이 직선 위에 위치하면 $|r|=1$
    - r가 음수이면 음의 상관관계가 존재
    - r가 양수이면 양의 상관관계가 존재
  - $|r| \simeq 0$이면 상관관계가 없다고 함
    - 어떤 관계도 존재하지 않는다는 것은 아님
  - $|r|$가 얼마 이상이어야 상관관계가 있다고 할 수 있는지?

    ⇨ "통계학의 이해II"

◉ 올림픽 개최 연도와 우승기록

○ 남자의 상관계수

$$S_{xx} = 103835088 - \frac{52940^2}{27} = 33473.19$$

$$S_{yy} = 2841.27 - \frac{276.76^2}{27} = 4.378$$

$$r_{xy} = \frac{-363.45}{\sqrt{33473.2}\ \sqrt{4.378}} = -0.949$$

○ 여자의 상관계수: -0.892

○ 연도와 우승기록 간에는 확실한 음의 상관관계가 있음

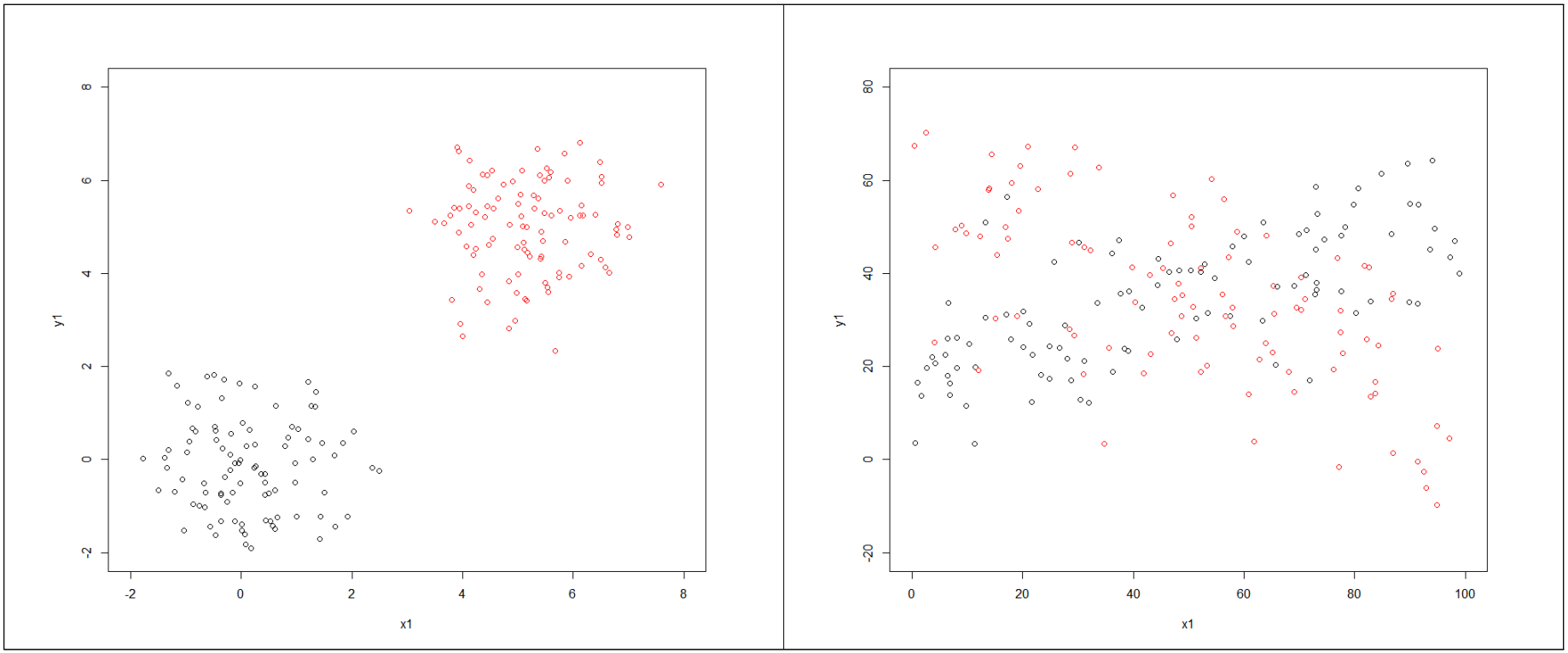
SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 상관관계 사용 시 주의할 점

- 두 변수 간에 직선관계가 있는지를 나타낼 뿐 인과관계를 나타내는 것은 아님
  - 예】휴대전화 보급률과 기대수명에 대한 상관계수
    - 매우 높은 양의 상관관계를 가짐
      ⇨ 기대수명을 늘리기 위해 휴대전화 보급을 늘려야 한다?
  - 잠복변수(lurking varaible): 두 변수에 영향을 주는 변수
    - 연도에 따라 보급률 증가, 기대수명 증가
      ⇨ 허위상관(spurious correlation)
    - 보급률과 기대수명에서 연도의 영향력을 제거하고 상관관계유도

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

• 통합된 그룹의 상관관계



SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

- 정리

  - 직선관계의 정도: 표본공분산, 표본상관계수
  - 표본상관계수: 표준화된 자료의 표본공분산
  - $|r|$이 1에 가까울수록 높은 상관관계
  - 주의할 점: 허위상관, 통합된 그룹의 상관관계