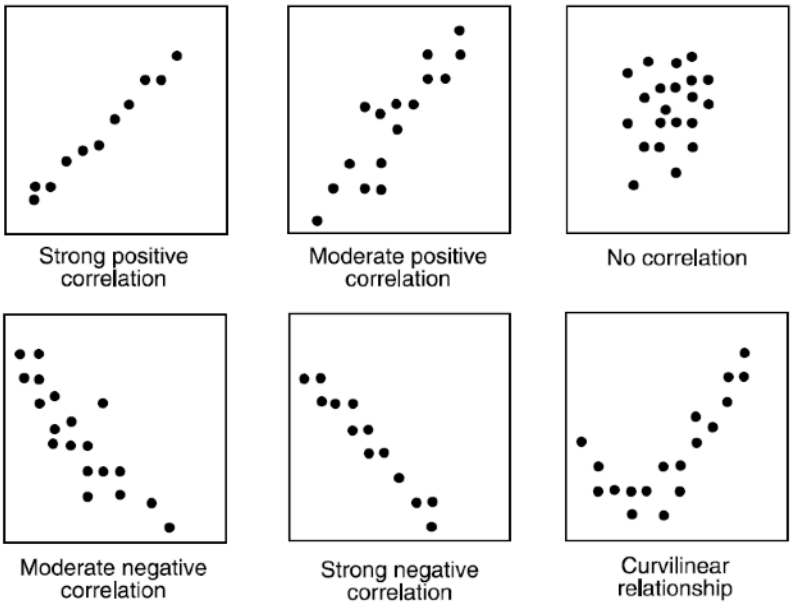
1. 개념

- 두 측정형 (적어도 순서형 범주형) 변수의 선형(직선)관계에 대한 척도
- 데이터 (x_i, y_i) 쌍으로 관측치를 활용
- 두 변수간의 관계를 시각적으로 표현하는 산점도는 두 변수 간의 함수 관계를 보여줌

2. 산점도 (SCATTER PLOT)

- 두 측정형 변수의 함수관계를 표현한 2차원 그래프
- 인과관계가 있다면 종속변수에 해당되는 변수를 y-축, 설명변수에 해당되는 변수를 x-축



(인터넷 그림 다운)

3. 상관계수 종류

(1) 피어슨 Pearson 상관계수

- 측정형 변수 간의 선형관계 척도
- (계산식) $\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2}\sqrt{E(Y^2) - E(Y)^2}}$. 모집단
- $r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$ 표본(2) 스피어맨 Spearman 순위 상관계수
- 순서형 변수 간의 선형 관계 척도

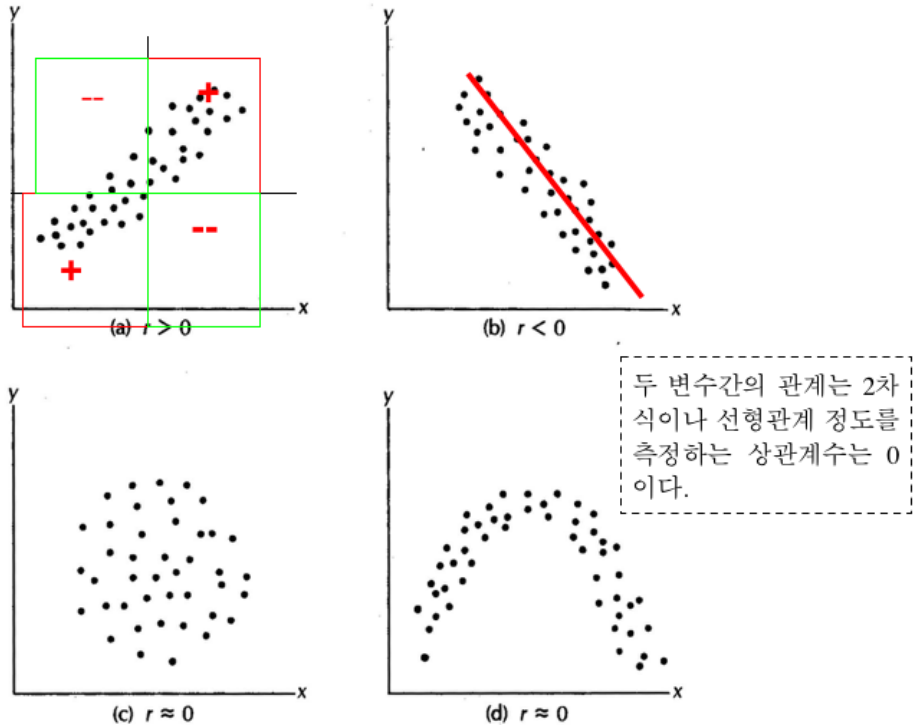
- (계산식) $\rho = \frac{6\sum d_i^2}{n(n^2-1)}$, where d_i 는 x_i 순위(x_i 관측치를 크기 순으로 정렬하였을 때 순위 rank)와 y_i 순위 차이

(3) Kendall Tau 순위 상관계수

- 순서형 변수 간의 선형 관계 척도
- concordant : 쌍의 관측치 값의 크기와 순위의 크기가 일치할 때
- (계산식) $\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)}$.

4. PEARSON 상관계수 추론

(계산식) $r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{E(X-E(X))E(Y-E(Y))}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$



두 변수간의 관계는 2차 식이나 선형관계 정도를 측정하는 상관계수는 0 이다.

(1) 상관관계 유의성 검정

- 귀무가설 : $H_0 : \rho = 0$ (두 측정형 변수의 상관관계는 유의하지 않음)
- 대립가설 : $H_0 : \rho \neq 0$ (두 측정형 변수의 상관관계는 유의하지 않음)

- 검정통계량 : $\frac{r-\rho_0(0)}{\sqrt{(1-r^2)/(n-2)}} \sim t(n-2)$

(2) $H_0 : \rho = \rho_0$ 검정

- 상관관계 유의성 검정이 아니라 임의의 상관계수와 동일한지 검정
- 활용 : 미국의 경우 부자 키의 상관계수는 0.65이다. 한국의 경우 미국과 부자의 키의 상관계수가 같다고 할 수 있나? 귀무가설 : $H_0 : \rho = 0.65$
- 검정통계량 : $TS = \frac{\frac{1}{2}\ln(\frac{1+r}{1-r}) - \frac{1}{2}\ln(\frac{1+\rho_0}{1-\rho_0})}{1/\sqrt{n-3}} \sim N(0,1)$

5. 두 독립집단 상관계수 차이 검정

- 귀무가설 $H_0 : \rho_x = \rho_y$
- 대립가설 $H_0 : \rho_x \neq \rho_y$
- 활용 : 한국 부자 키의 상관계수와 미국 부자 키의 상관계수는 동일한가?
- 검정통계량 :

$$z(x) = 0.5 \ln \frac{1+r_x}{1-r_x}, z(y) = 0.5 \ln \frac{1+r_y}{1-r_y}$$

$$z = \frac{z(x) - z(y)}{\sqrt{1/(n_x - 3) + 1/(n_y - 3)}} \sim N(0,1)$$

6. 회귀계수와 관계

- 단순 회귀모형 $y_i = \alpha + \beta x_i + e_i$ 에서 회귀계수 OLS 추정치 $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- 상관계수와 회귀계수 관계식 $\hat{\beta} = \sqrt{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \times r$: 부호가 동일하며 비례관계
- 상관계수 유의성 검정과 회귀계수 기울기 유의성 검정은 동일하며 $t(n-2)$ 샘플링분포

7. 단순회귀모형 결정계수와 관계

(a) 결정계수

- 단순 회귀모형에서 결정계수 Determination Coefficient $R^2 = \frac{SSR}{SST}$, $0 < R^2 < 100(\%)$
- 총변동 중 회귀변동이 차지하는 비율 : 모형의 적합 정도를 나타냄

(b) 관계

- 상관계수의 제곱 = 결정계수 $r^2 = R^2$

$$
\begin{aligned}
r(Y,\hat{Y}) &= \frac{\sum_i (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i [(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + (\hat{Y}_i - \bar{Y})^2]}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \cdot \sum_i (\hat{Y}_i - \bar{Y})^2}} \\
&= \sqrt{\frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2}}.
\end{aligned}
$$

- (Wikipedia) => $r(Y,\hat{Y})^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$

8. 상관계수 해석의 유의사항

- 양의 부호 : 한 변수 값이 커지면(작아지면) 다른 변수 값도 커진다(작아진다)
- 음의 부호 : 한 변수 값이 커지면(작아지면) 다른 변수 값도 작아진다(커진다)
- 상관관계 유의성은 크기로 결정하는 것이 아니라 검정 결과의 “유의확률”의 크기에 의해 판단
- 상관계수의 값의 크기와 상관관계 유의성은 비례하는 것은 아님 - 왜냐하면 측정변수의 관측값이 충분히 연속형이 아닌 경우 (예를 들면 일주일 교통사고 건수처럼 0. 1, 2, .., 7이면 상관계수 값는 낮을 수 있음)
- 데이터 개수가 많아지면 상관계수 값의 크기는 무조건 커진다 (가장 큰 단점)
- 두 측정형 변수 : 상관계수 0.7이상(little correlated), 0.8이상(correlated), 0.9이상(highly)

(실습) ❑BABE.csv

베이브 루즈가 선수 시절 프로선수들의 OBA(Opponents' Batting Average), EBP(extra base power)=OBP(on-base percentage) + SLG(slugging percentage).

(a) 산점도를 그리고 해석하시오.

(b) 상관계수를 계산하고 해석하시오.

It is calculated as: $BAA = \frac{H}{BF - BB - HBP - SH - SF - CINT}$

for which:

- BF is the number of batters faced by the pitcher
- BB is the number of base on balls
- HBP is the number of hit batsmen
- SH is the number of sacrifice hits
- SF is the number of sacrifice flies
- CINT is the number of catcher's interference

$$SLG = \frac{(1B) + (2 \times 2B) + (3 \times 3B) + (4 \times HR)}{AB}$$

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}$$

where

- *H* = Hits
- *BB* = Bases on Balls (Walks)
- *HBP* = Hit By Pitch
- *AB* = At bats
- *SF* = Sacrifice Flies

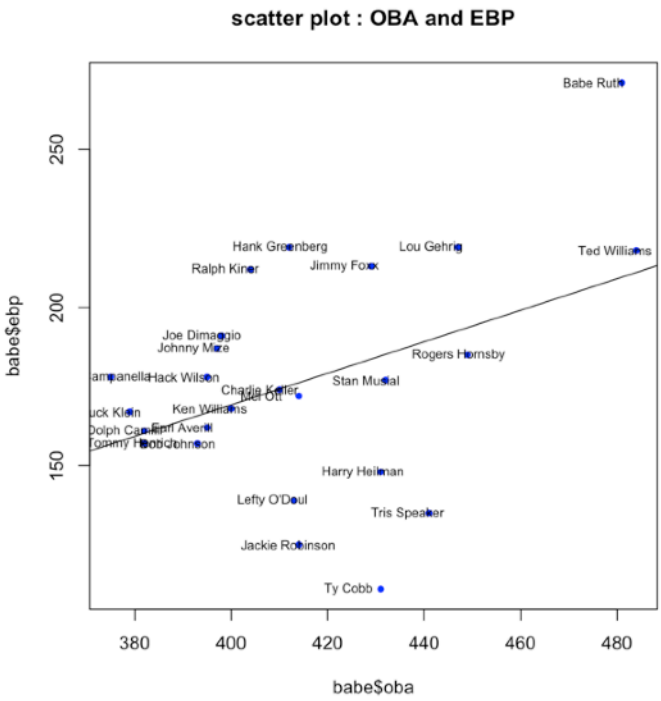
```
babe<-read.csv("babe_homerun.csv")
names(babe)
plot(babe$oba,babe$ebp, col="blue", pch=20,
     main="scatter plot : OBA and EBP")     #scatter plot
text(babe$oba,babe$ebp,labels=babe$player, cex= 0.7)   #id labels
cor.test(babe$oba,babe$ebp,method="pearson") #피어슨 상관분석
abline(lm(babe$ebp~babe$oba))
```

```
> cor.test(babe$oba,babe$ebp)

        Pearson's product-moment correlation

data:  babe$oba and babe$ebp
t = 2.1764, df = 23, p-value = 0.04005
alternative hypothesis: true correlation is r
95 percent confidence interval:
 0.02165601 0.69491093
sample estimates:
      cor
0.4132507
```

선수의 파워와 정교함은 양의 상관관계 0.41(유의확률 0.04) 유의하지만 상관관계 정도는 악하다.



약한 양의 직선(상관) 관계 보임

- Babe - 영향치 & 이상치
- TED - 영향치
- 아래 선수군 - 이상치

```
cor.test(babe[c(-1,-2),]$oba,babe[c(-1,-2),]$ebp,method="pearson") #babe, ted 삭제
```

```
data:  babe[c(-1, -2), ]$oba and babe[c(-1, -2), ]$ebp
t = 0.12183, df = 21, p-value = 0.9042
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.389897  0.434024
sample estimates:
       cor
0.02657651
```

영향치 2개 삭제 - 상관계수 낮아지고 유의성 상실, 두 변수는 상관관계가 존재하지 않는다.