제6강
다중변수 자료의 탐색

Section 01

산점도

Section 02

상관 분석

2. 상관분석

1. 상관분석과 상관계수

- 자동차의 중량이 커지면 연비는 감소하는 추세
- 추세의 모양이 선(線, line) 모양이어서 중량과 연비는 '선형적 관계'에 있다고 표현
- 선형적 관계라고 해도 강한 선형적 관계가 있고 약한 선형적 관계도 있음
- 상관분석(correlation analysis) : 얼마나 선형성을 보이는지 수치상으로 나타낼 수 있는 방법

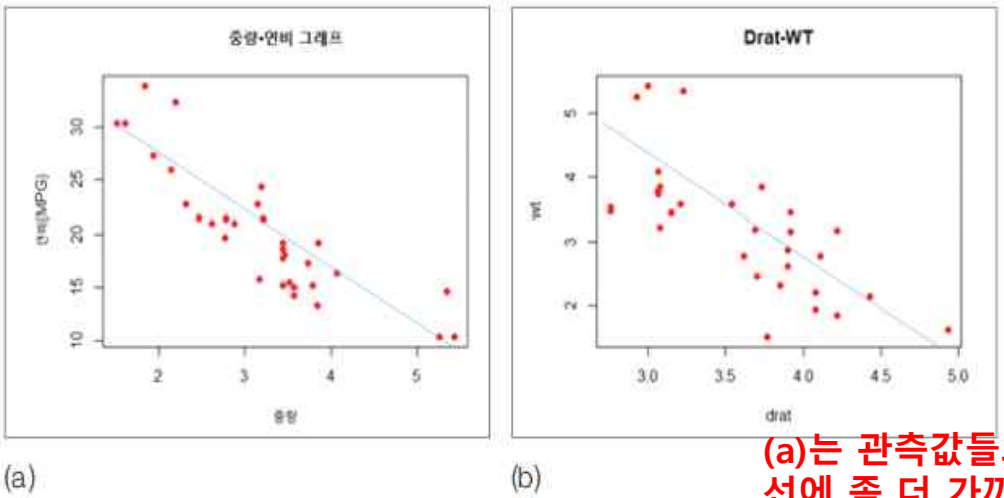


그림 6-4 선형적 관계에 있는 두 변수

(a)는 관측값들의 분포가 (b)에 비하여 선에 좀 더 가까운 것을 알 수가 있다. 즉, 강한 선형적 관계라고 볼 수 있다.

2. 상관분석

- **피어슨 상관계수(Pearson's correlation coerricient) : 선형성의 정도를 나타내는 척도로 사용됨**

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

- **-1 ≤ r ≤ 1**
- **r > 0 : 양의 상관관계(x가 증가하면 y도 증가)**
- **r < 0 : 음의 상관관계(x가 증가하면 y는 감소)**
- **r이 1이나 –1에 가까울수록 x, y의 상관성이 높음**

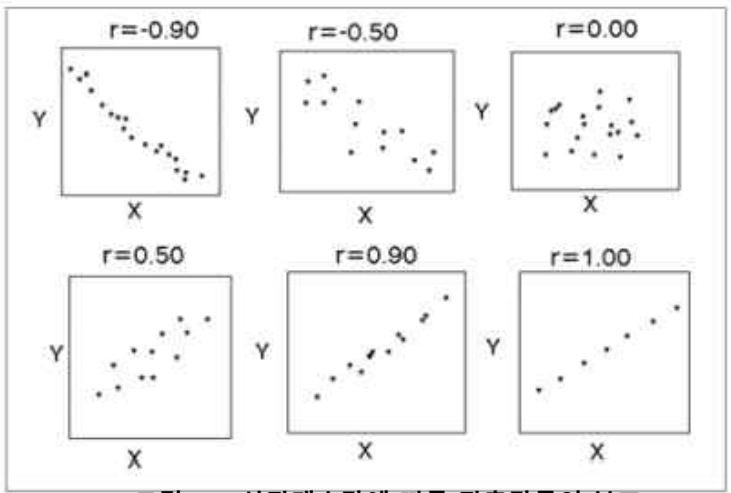


그림 6-5 상관계수값에 따른 관측값들의 분포

2. 상관분석

2. R을 이용한 상관계수의 계산

- 음주정도와 혈중알콜농도가 상관성 조사

beers	5	2	9	8	3	7	3	5	3	5
bal	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06	0.02	0.05

코드 6-4

```
beers = c(5,2,9,8,3,7,3,5,3,5)                  # 자료 입력
bal <- c(0.1,0.03,0.19,0.12,0.04,0.0095,0.07,   # 자료 입력
         0.06,0.02,0.05)
tbl <- data.frame(beers,bal)                    # 데이터프레임 생성
tbl
plot(bal~beers, data=tbl)                       # 산점도
res <- lm(bal~beers,data=tbl)                   # 회귀식 도출
abline(res)                                     # 회귀선 그리기
cor(beers,bal)                                  # 상관계수 계산
```

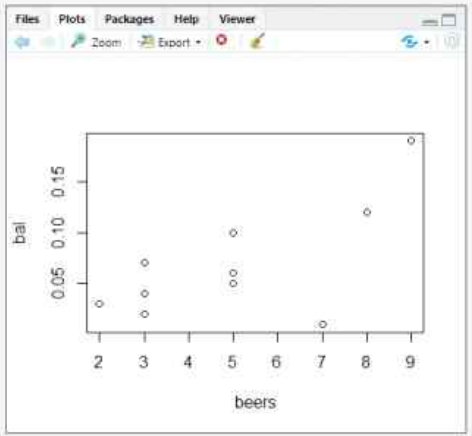
lm()함수는 두 변수의 선형 관계를 가장 잘 나타낼수 있는 선의 식을 자동으로 찾는 역할을 한다.
여기서, 두 변수의 선형 관계를 가장 잘 나타내는 선의 식을 회귀식이라고 한다.
abline()함수는 회귀식을 이용하여 산점도 위에 회귀선을 그리는 함수이다.
하여, 회귀선은 관측값들의 추세를 가장 잘 나타낼 수 있는 선이다.

2. 상관분석

```
> beers <- c(5,2,9,8,3,7,3,5,3,5)                # 자료 입력
> bal <- c(0.1,0.03,0.19,0.12,0.04,0.0095,0.07,  # 자료 입력
+        0.06,0.02,0.05)
> tbl <- data.frame(beers,bal)                   # 데이터프레임 생성
> tbl
   beers    bal
1      5 0.1000
2      2 0.0300
3      9 0.1900
4      8 0.1200
5      3 0.0400
6      7 0.0095
7      3 0.0700
8      5 0.0600
9      3 0.0200
10     5 0.0500
```

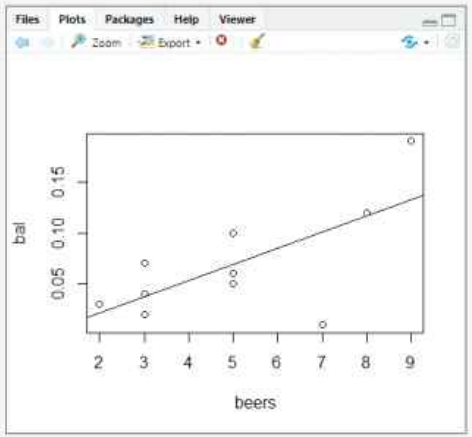
2. 상관분석

```
> plot(bal~beers,data=tbl)                    # 산점도
```



2. 상관분석

```
> res=lm(bal~beers,data=tbl)          # 회귀식 도출
> abline(res)                         # 회귀선 그리기
```




```
> cor(beers,bal)                      # 상관계수 계산
[1] 0.6797025
```

cor()함수는 상관계수를 구하는역할을 한다. 상관계수는 어느 정도 되어야 두 변수가 상관성이 있을까? 이와 관련해서 정해진 것은 딱히 없다.
하지만, 상관계수 값이 0.5보다 크거나 -0.5보다 작으면 두 변수의 상관성은 높다고 보는 것이 일반적이다.

2. 상관분석

코드 6-5

```
cor(iris[,1:4])                    # 4개 변수 간 상관성 분석
```

```
> cor(iris[,1:4])                  # 4개 변수 간 상관성 분석
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

결과를 분석해보면, 4개의 변수가 x축, y축방향으로 나열되어 있고, 두 변수가 만나는 지점에 상관계수가 표시 되어 있다.
Petal.Length와 Sepal.Width의 상관계수 값은 -0.428정도이며, 여기서 가장 상관성이 높은 변수들은 Petal.Length와 Petal.Width이며 값이 0.962정도로 거의 1에 가까운 결과를 볼 수가 있다.
즉, 그만큼 상관성이 높다는 것을 의미한다.

감사합니다.