R을 활용한 연관성 분석

호서대학교 빅데이터경영공학부 연규필

양적자료간 연관성 분석

1. 산점도
2. 상관분석
3. 편상관계수

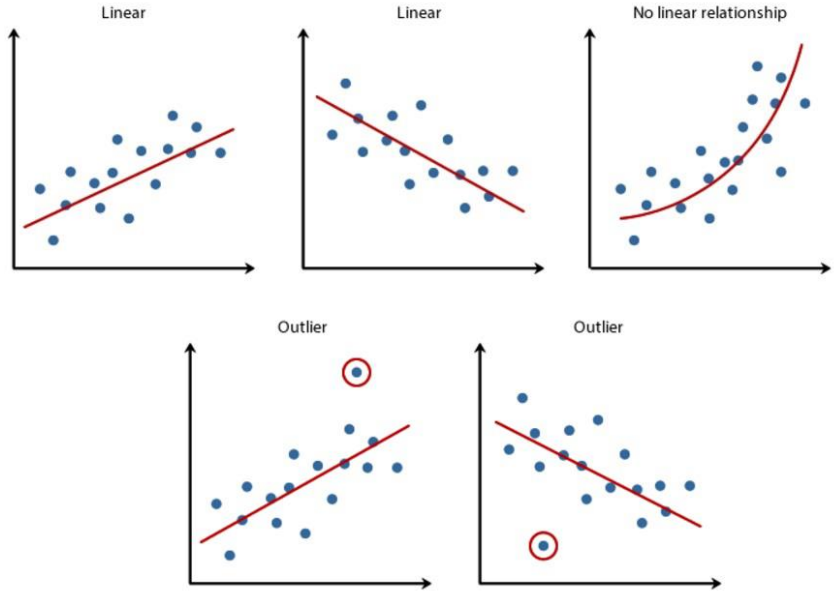
1. 산점도

1.1 산점도

■ 산점도

▪ 산점도 (scatter plot)

- n개의 개체에 대해 이변량 관측벡터 $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$의 n쌍의 관측값이 주어질 때, 이들 각 개체에 대한 관측값들을 2차원 공간에서 한 점으로 표현한 그래프를 산점도(scatter plot)라 함.
- 산점도를 통해서 두 변수의 관계가 선형(linear)인지 비선형(non-linear)인지 탐색하고 이상치를 찾아낼 수 있으며 변수관계를 시각적으로 표현할 수 있음



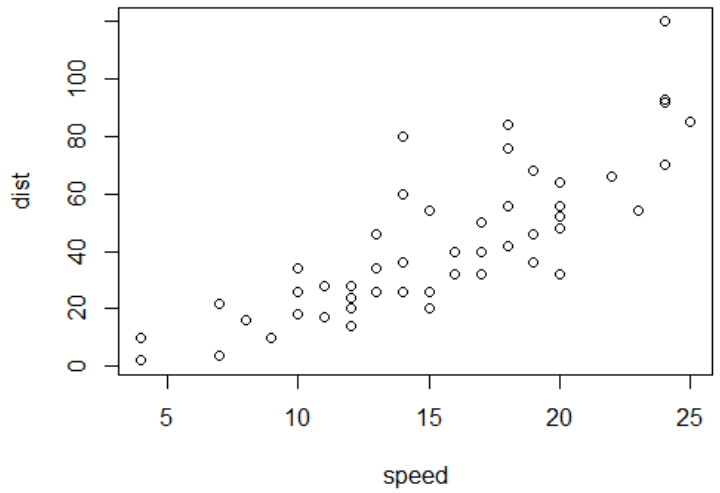
1.1 산점도

■ 산점도

▪ 예제 cars 데이터

- 차량의 속도(speed)와 정지까지의 거리(stopping distance)에 대한 실험 데이터 (1920년)
  - speed: 자동차의 주행속도, dist : 브레이크를 밟았을 때의 제동 거리

```
> str(cars)
'data.frame':           50 obs. of  2 variables:
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
> attach(cars)
> plot(speed,dist)
```



1.1 산점도

■ 산점도

▪ 예제 mtcars 데이터

- 1974년 Moto US magazine의 1973-1974년 32개 자동차 모형에 대한 자료

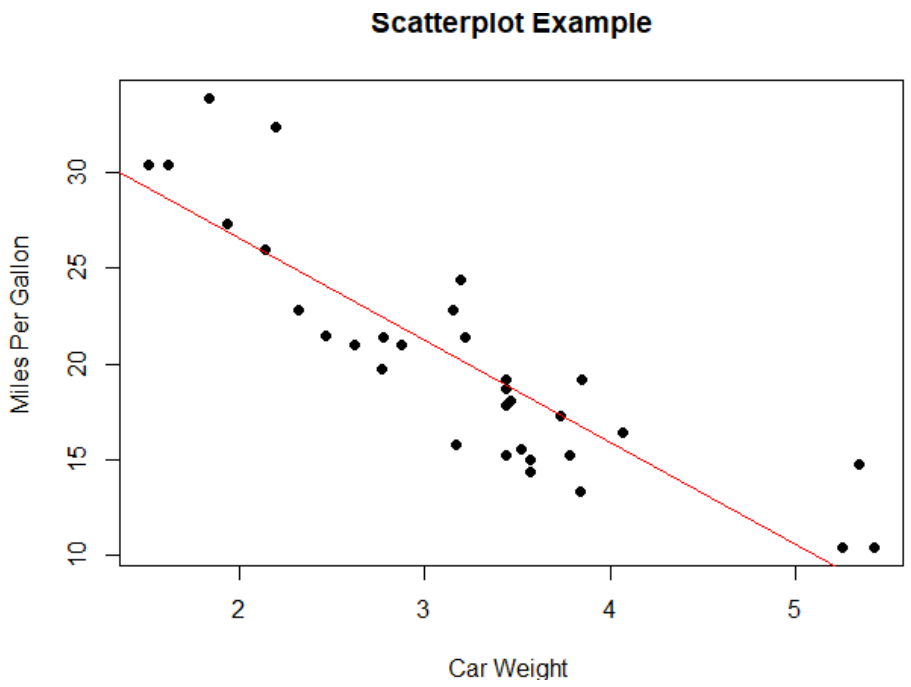
[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

1.1 산점도

■ 산점도

- 예제 mtcars 데이터

```
> attach(mtcars)
> plot(wt, mpg, main = "Scatterplot Example", xlab = "Car Weight ",
       ylab = "Miles Per Gallon ", pch = 19)
> abline(lm(mpg ~ wt), col = "red")
```

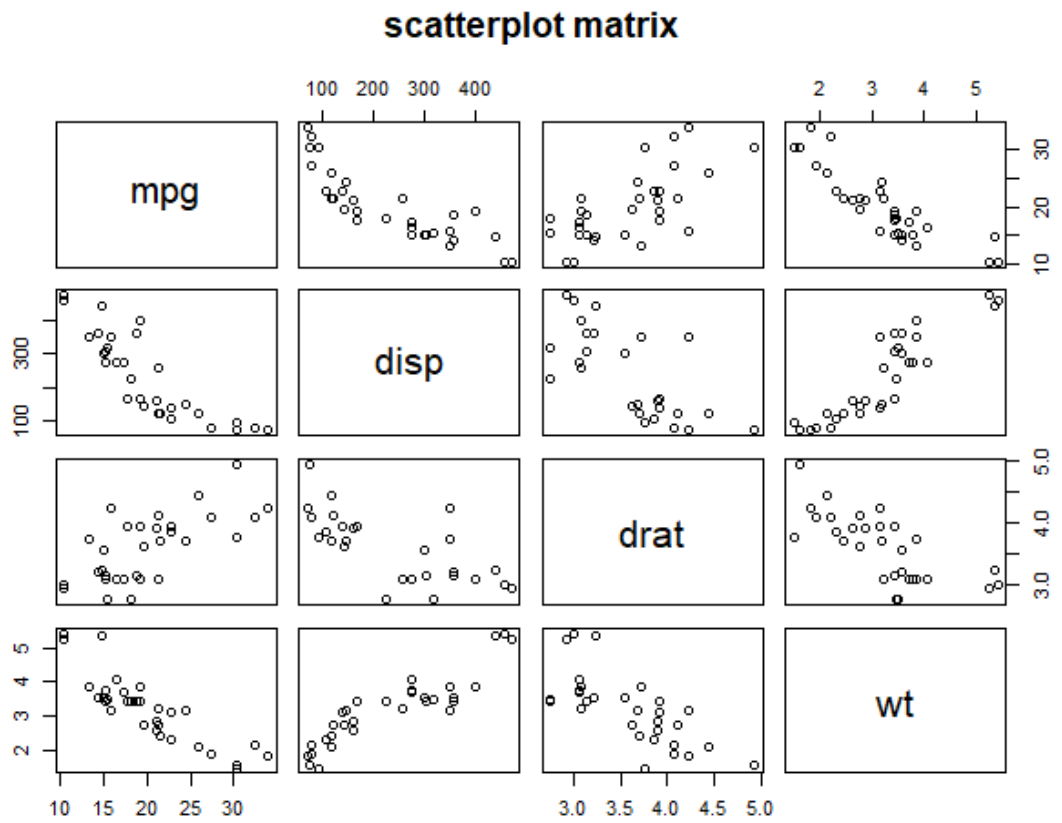


1.1 산점도

■ 산점도 행렬 (scatter plot matrix)

- 예제 mtcars 데이터

```
> pairs(~mpg + disp + drat + wt, main = "scatterplot matrix")
```



1.1 산점도

■ 산점도 행렬 (scatter plot matrix)

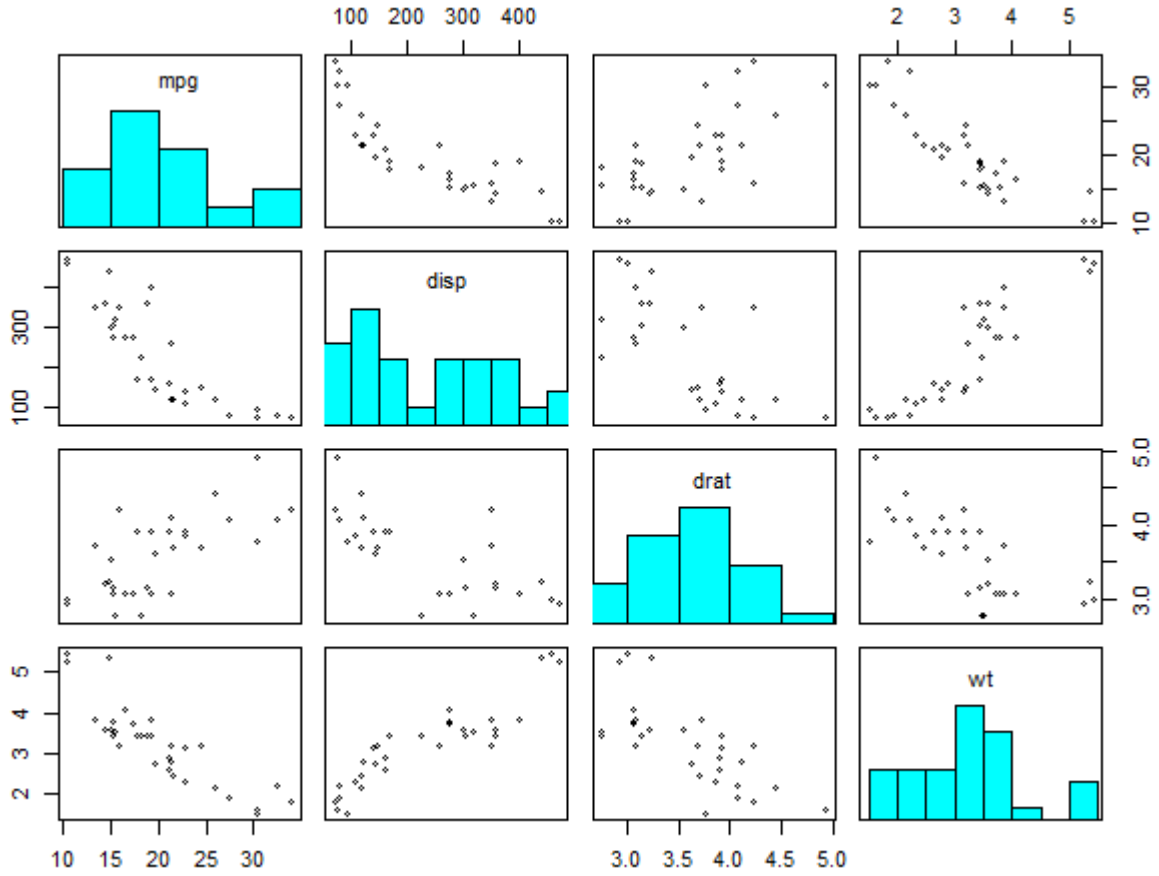
- 예제 mtcars 데이터

```
> panel.hist <- function(x, ...) {
  usr <- par("usr")
  on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5))
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks
  nB <- length(breaks)
  y <- h$counts
  y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan",
       ...)
}
> pairs(~mpg + disp + drat + wt, cex = 0.7, bg = "light blue",
      diag.panel = panel.hist, cex.labels = 1, font.labels = 1)
```

1.1 산점도

■ 산점도 행렬 (scatter plot matrix)

- 예제 mtcars 데이터



2. 상관분석

2.1 상관계수

■ 표본상관계수

- 이변량 (bivariate) 데이터 $(X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)$ 가 주어질 때, 두 확률변수 사이의 관련성은 상관계수를 통해 나타낸다.
- 모집단 상관계수

$$\rho = cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- 표본 상관계수(Pearson correlation coefficient)

$$r = \hat{\rho} = \frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

여기서

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i, \qquad s_{XY} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}),\ s_{XX} = \sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2},\ s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2},\ s_{XY} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$$

2.1 상관계수

■ 표본상관계수

- **상관계수의 범위 :** $-1 \le \rho \le 1$, **표본상관계수의 범위 :** $-1 \le r \le 1$
  - 직선관계를 설명, 부호는 방향을 나타내고, 절대값 크기는 정도를 나타낸다.
  - r이 1에 가까울수록 양의 상관관계가 강하고, r이 -1에 가까울수록 음의 상관관계가 강함을 의미함.
- **상관계수의 검정**
  - 검정통계량

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$$

  - **기각역**

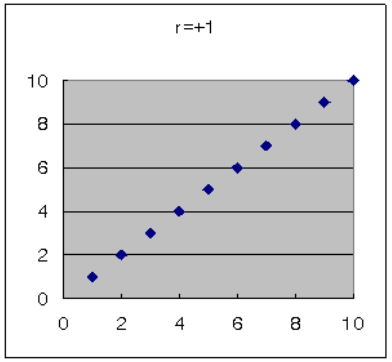
귀무가설	대립가설	유의수준 α 인 기각역
$H_0 : \rho \le 0$	$H_0 : \rho \le 0$	$T \ge t(n-2, \alpha)$
$H_0 : \rho \ge 0$	$H_0 : \rho \ge 0$	$T \le -t(n-2, \alpha)$
$H_0 : \rho = 0$	$H_0 : \rho = 0$	$\lvert T \rvert \ge t(n-2, \alpha/2)$

  - cor 함수를 이용하여 상관계수를 구하고 cor.test 함수를 이용하여 상관성검정을 함

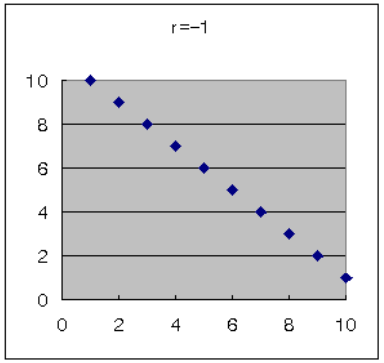
2.1 상관계수

■ 표본상관계수

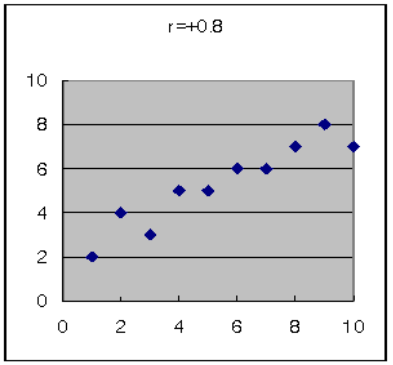
① r=1



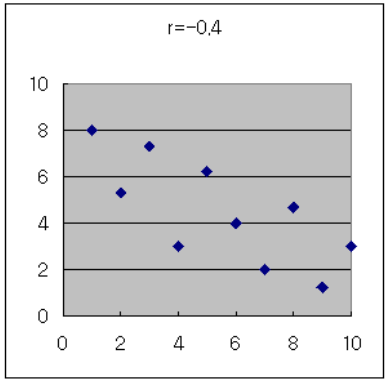
② r=−1



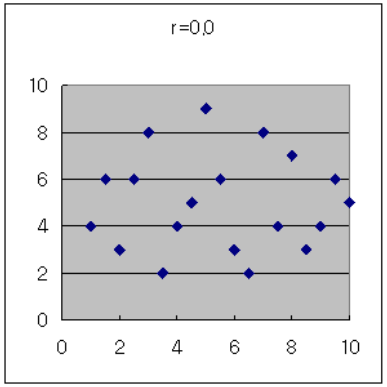
③ r=0.8



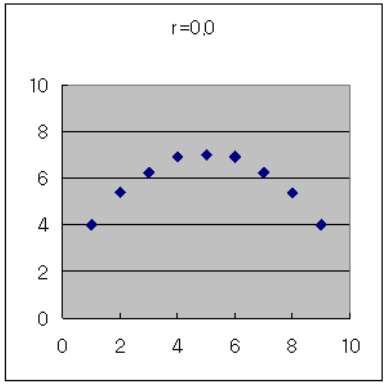
④ r=−0.4



⑤ r=0



⑥ r=0



2.1 상관계수

■ 상관계수의 종류

- **Pearson 상관계수 (Pearson's product-moment correlation coefficient)**
  - 가장 보편적인 상관계수로 간격/비율 척도로 측정된 자료들 간의 선형관계를 나타냄

    (예) $y = x^2, \quad x = -20, -19, \cdots, 19, 20$

- **Spearman 상관계수**
  - 서열 척도로 측정된 자료들 간의 상관 관계

- **Kendall's Tau 상관계수**
  - 서열 척도로 측정된 자료들 간의 상관 관계

- Spearman 상관계수와 Kendall's Tau 상관계수
  - 한 변수가 증가할 때 다른 변수가 증가하는 지의 관계를 봄으로 이 관계가 선형일 필요는 없음.
  - 증가정도가 계산에 반영되지 않아 연속형 변수라도 이상치가 있는 경우 유용하게 사용할 수 있음

    (예) $(x, y) = (0,1), (10,100), (101,1000), (102,5000)$

2.1 상관계수

■ 표본상관계수

▪ cor 함수

함수		Arguments
cor(x, y = NULL, method = c("pearson", "kendall", "spearman")),	x	• 데이터 값의 숫자 벡터
	y	• 데이터 값의 숫자 벡터
	method	• "pearson"(default), "kendall", or "spearman"를 나타내는 문자

```
> attach(cars)
> cor(speed, dist)
[1] 0.8068949
> cor(speed, dist, method = "pearson")
[1] 0.8068949
> cor(speed, dist, method = "kendall")
[1] 0.6689901
> cor(speed, dist, method = "spearman")
[1] 0.8303568
```

2.1 상관계수

■ 상관계수 유의성 검정

▪ cor.test 함수

함수		Arguments
cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)	x	• 데이터 값의 숫자 벡터
	y	• 데이터 값의 숫자 벡터
	alternative	• "two.sided"(default), "greater" or "less"의 양측, 단측
	method	• "pearson"(default), "kendall", or "spearman"의 문자
	exact	• exact p-value
	conf.level	• 신뢰수준
	continuity	• 연속성 보정
cor.test(formula, data, subset, na.action, ...)	formula	• formula (lhs ~ rhs) 형태. rhs는 집단을 나타내는 2 수준의 factor임
	data	• 메트릭스 또는 데이터프레임
	subset	• 부분집합을 나타내는 벡터
	na.action	• NA를 포함할 때 사용하는 함수

2.1 상관계수

■ 상관계수 유의성 검정

- cor.test 함수

```
> cor.test(speed, dist, method = "pearson")

        Pearson's product-moment correlation

data:  speed and dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```

상관계수는 0.8068로 양의 상관성이 매우 강하고 상관성 검정결과 p-값=1.49e-12으로 유의수준 5%에서 매우 유의하여 귀무가설을 기각함. 즉 속도와 거기 간의 상관관계는 매우 유의하다.

2.2 상관계수의 해석

■ 표본상관 해석 시 유의사항

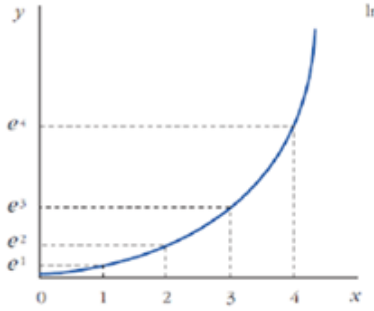
- **상관계수는 선형관계의 상관성만을 나타낸다.**
  - **상관계수 $r = 0$ 은 선형관계가 아님을 의미하지 (X_i, Y_i)가 관련성이 없음을 의미하는 것은 아니다.**
- **상관관계가 유용하지 않은 경우**

이탈값이 있는 경우

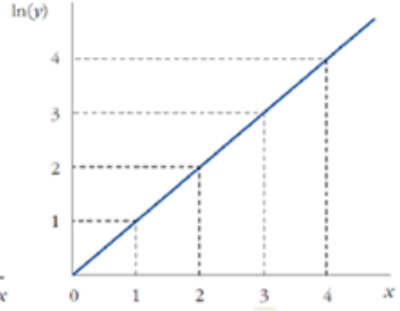
비선형회귀

- **비선형 관계는 변수변환을 통하여 선형관계로 변환한다.**

원래의 비선형 관계



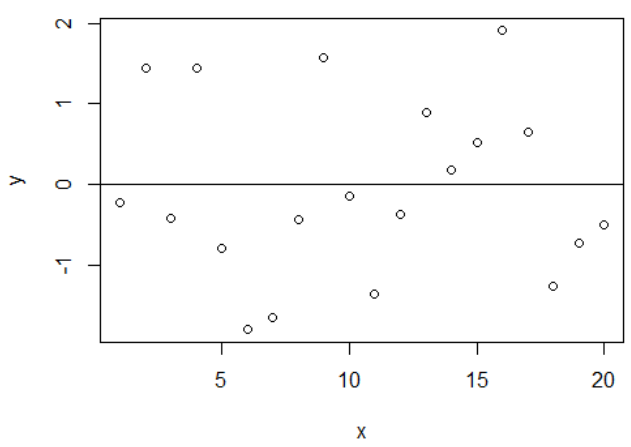
선형 관계로 변환



2.2 상관계수의 해석

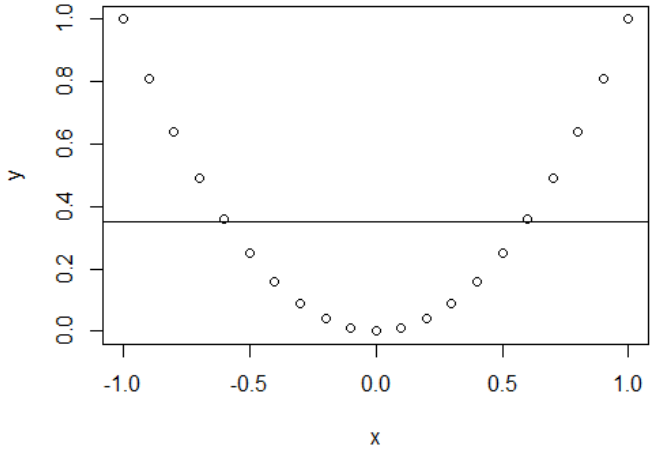
■ 표본상관 해석 시 유의사항

```
> x=seq(1,20,by=1)
> y=runif(20,min=-2, max=2)
> cor(x,y)
[1] -0.009058125
> plot(x,y)
> abline(h=0)
```



$r = -0.0091$

```
> x=seq(-1.0,1.0,by=0.1)
> y=x^2
> cor(x,y)
[1] 1.216307e-16
> plot(x,y)
> abline(h=0.35)
```



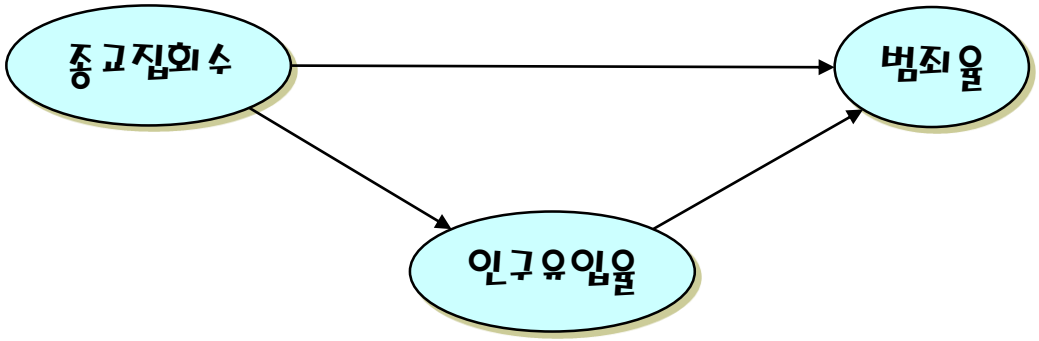
$r = 0$

3. 편상관계수

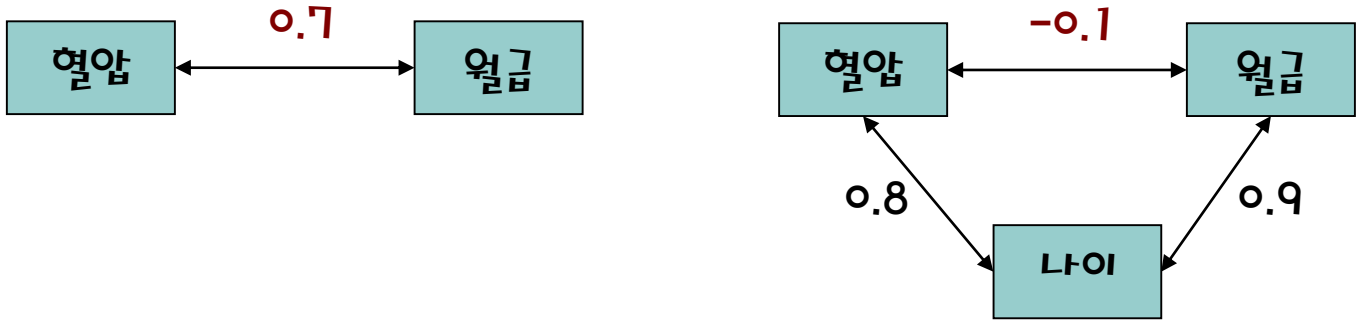
3.1 편상관계수

■ 표본상관 해석의 제약성: 의사상관

- 실제적인 연관관계가 없음에도 불구하고 상관계수가 크게 나타나는 경우.
- 제 3의 변수에 의하여 상관관계가 나타나는 경우.



- **편상관계수 (partial correlation coefficient)**



3.1 편상관계수

■ 사례 분석

- Satis 데이터

ID	Age	Satis1	Satis2	ID	Age	Satis1	Satis2
1	28	0	70	11	31	40	75
2	23	0	55	12	33	50	80
3	26	5	65	13	39	55	95
4	27	5	65	14	36	60	90
5	25	10	60	15	30	65	75
6	26	20	65	16	36	65	90
7	29	25	70	17	32	80	80

```
Satis = read.csv("data/Satis.csv",header=TRUE)
cor(Satis[c("Age","Satis1","Satis2")])
cor.test(Satis$Satis1,Satis$Satis2)
```

```
> cor(Satis[c("Age","Satis1","Satis2")])
              Age    Satis1    Satis2
Age    1.0000000 0.6982138 0.9934652
Satis1 0.6982138 1.0000000 0.7030501
Satis2 0.9934652 0.7030501 1.0000000
> cor.test(Satis$Satis1,Satis$Satis2)

        Pearson's product-moment correlation

data:  Satis$Satis1 and Satis$Satis2
t = 4.1944, df = 18, p-value = 0.000545
```

3.1 편상관계수

■ 사례 분석

- 편상관계수의 계산

```
install.packages("ggm")
library(ggm)
pcor(c("Satis1", "Satis2", "Age"),var(Satis))
Satis.pcor = pcor(c("Satis1", "Satis2", "Age"),var(Satis))
pcor.test(Satis.pcor,1,20)
```

```
> pcor(c("Satis1", "Satis2", "Age"),var(Satis))
[1] 0.1150319
> Satis.pcor = pcor(c("Satis1", "Satis2", "Age"),var(Satis))
> pcor.test(Satis.pcor,1,20)
$tval
[1] 0.477458

$df
[1] 17

$pvalue
[1] 0.6391169
```

3.1 편상관계수

■ 사례 분석

- 연령대별 산점도

```
Satis$Agegroup[Satis$Age<30]=1
Satis$Agegroup[Satis$Age>=30]=2
Satis$Agegroup = factor(Satis$Agegroup,labels=c("20대","30대"))
library(car)
scatterplot(Satis2~Satis1,data=Satis,cex=2,regLine=list(col="green"),lwd=2,boxplots=FALSE,smooth=FALSE)
scatterplot(Satis2~Satis1,data=Satis,groups=Satis$Agegroup,cex=2,lwd=2,boxplots=FALSE,smooth=FALSE)
```