변수유형에 따른 분석기법

		설명변수 (분석축)			
		한 그룹과 특정 숫자와의 비교	두 그룹의 비교	셋 이상 그룹의 비교	양적변수 (연속값)의 크기로 비교
반응변수 (분석하고 싶은 것)	양적변수 (연속값 등)	**One-sample T-test**	**Two-sample T-test Paired T-test**	**분산분석 (ANOVA)**	**회귀분석**
	질적변수 (Yes/No)	**One- proportion Z-test**	**Two-proportion Z-test**	**분할표 카이제곱 검정**	**로지스틱 회귀 분석**

상관분석

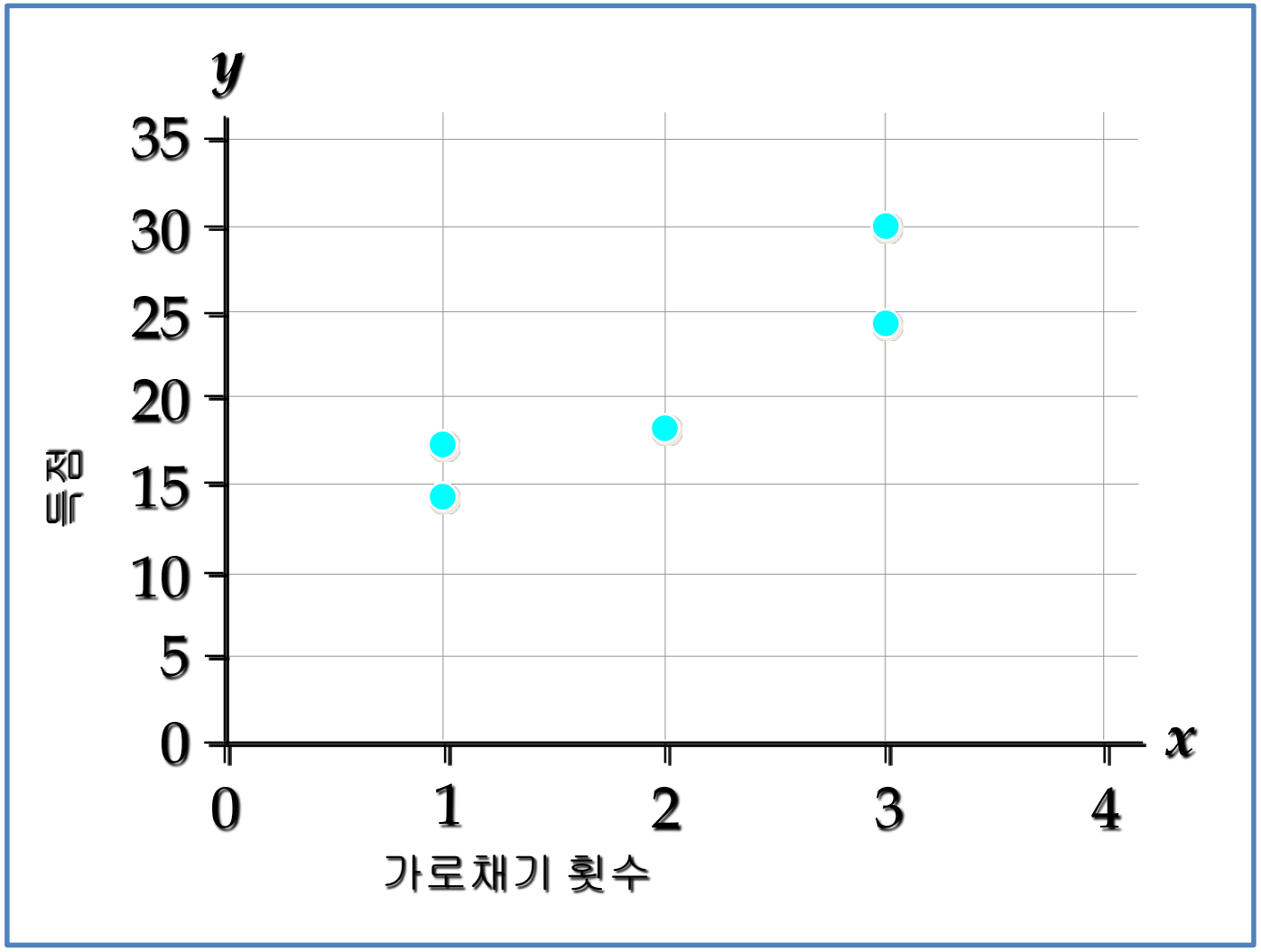
예 : Panthers Football Team

- 산점도

  Panthers football team은 가로채기 횟수와 득점과의 관계에 대하여 조사하고 싶어 한다.

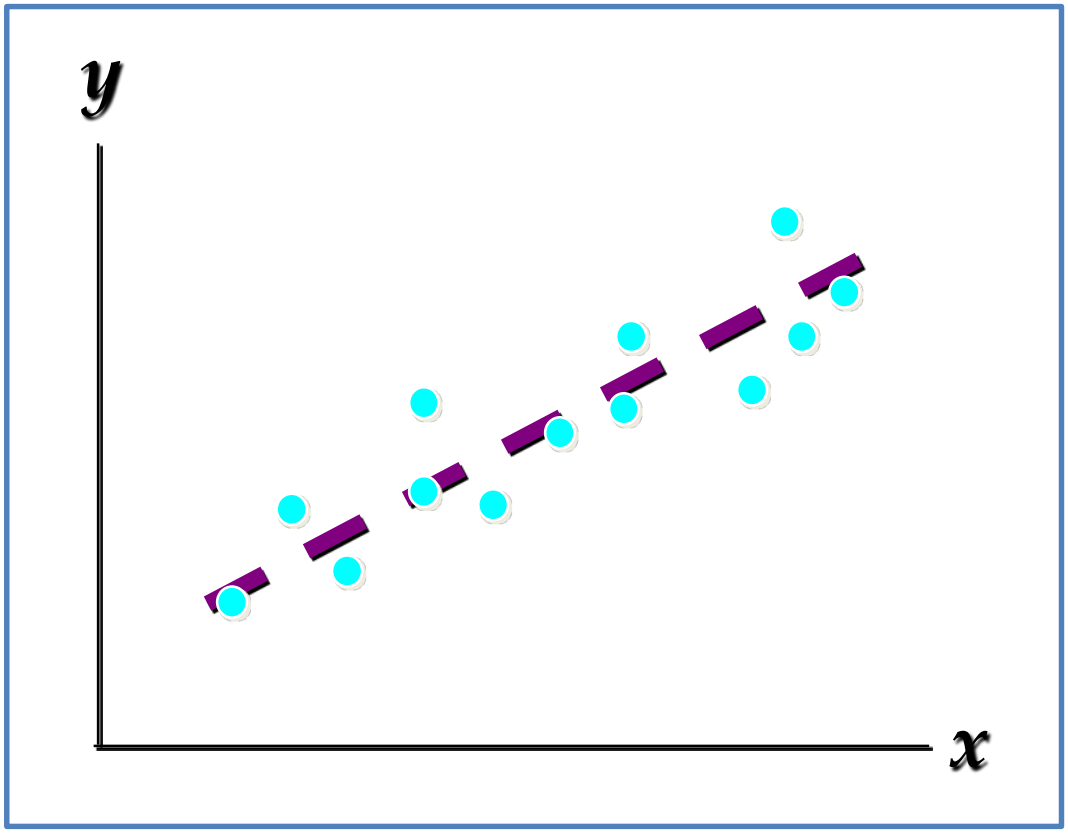
x = 가로채기 횟수	y = 득점
1	14
3	24
2	18
1	17
3	30

산점도



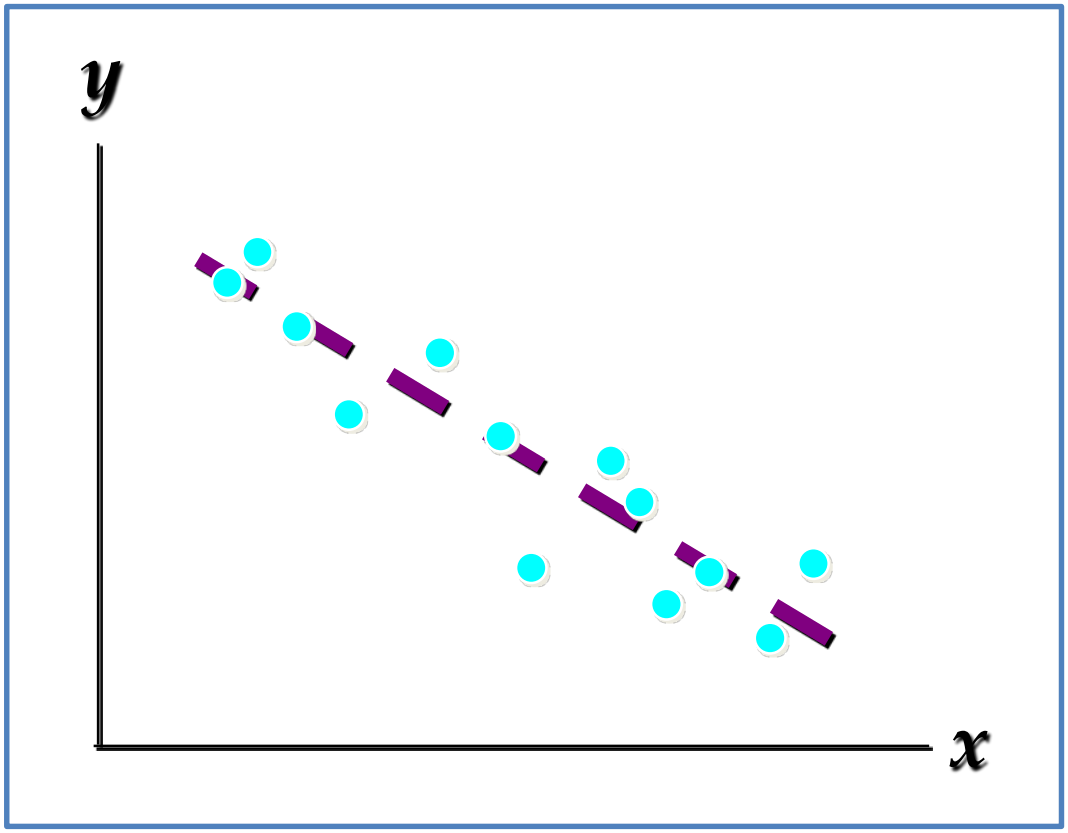
산점도

- 정의 관계



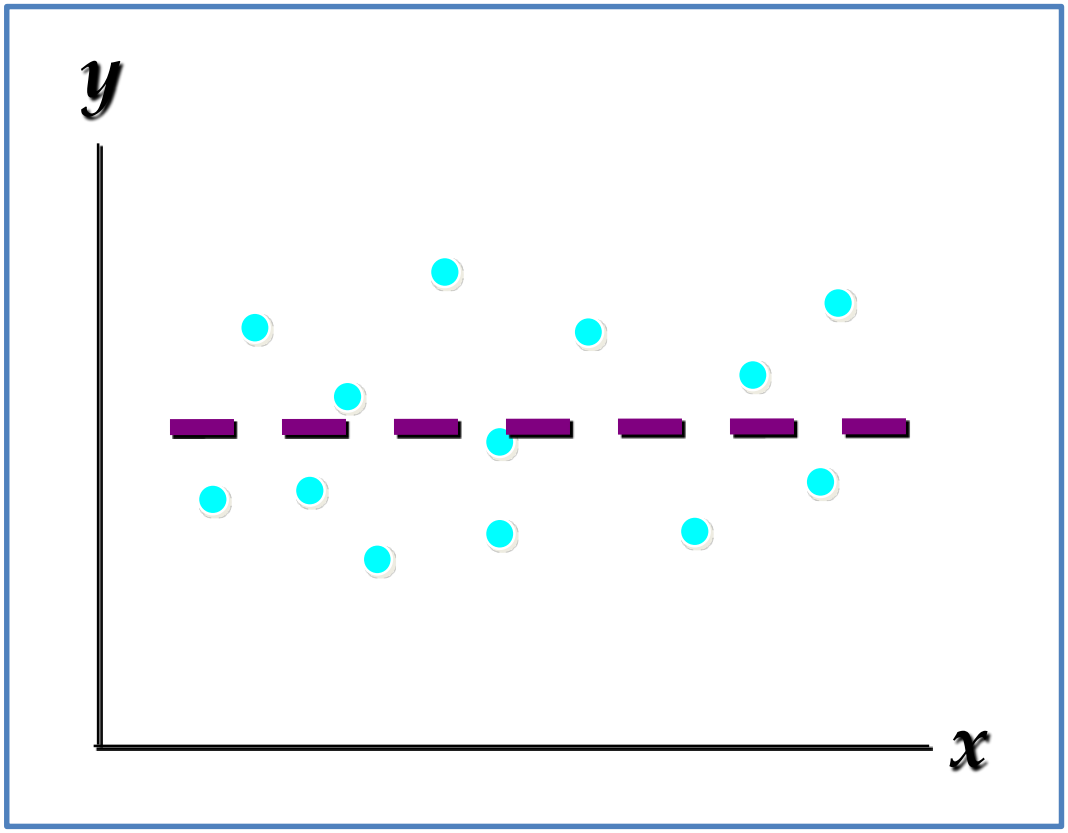
산점도

- 역의 관계



산점도

- 뚜렷한 관계가 없는 경우



Pearson 상관계수

- 공분산 $cov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
  - 두 변수가 같은 방향으로 움직이는 정도를 측정
  - 측정단위에 영향을 받는다 (kg vs g, km vs mile)
- 상관계수 $corr(x, y) = \frac{\mathrm{cov(x,y)}}{sd(x)sd(y)}$
  - 표준편차로 나누어 주어 언제가 -1과 1 사이의 값

Pearson 상관계수의 특징

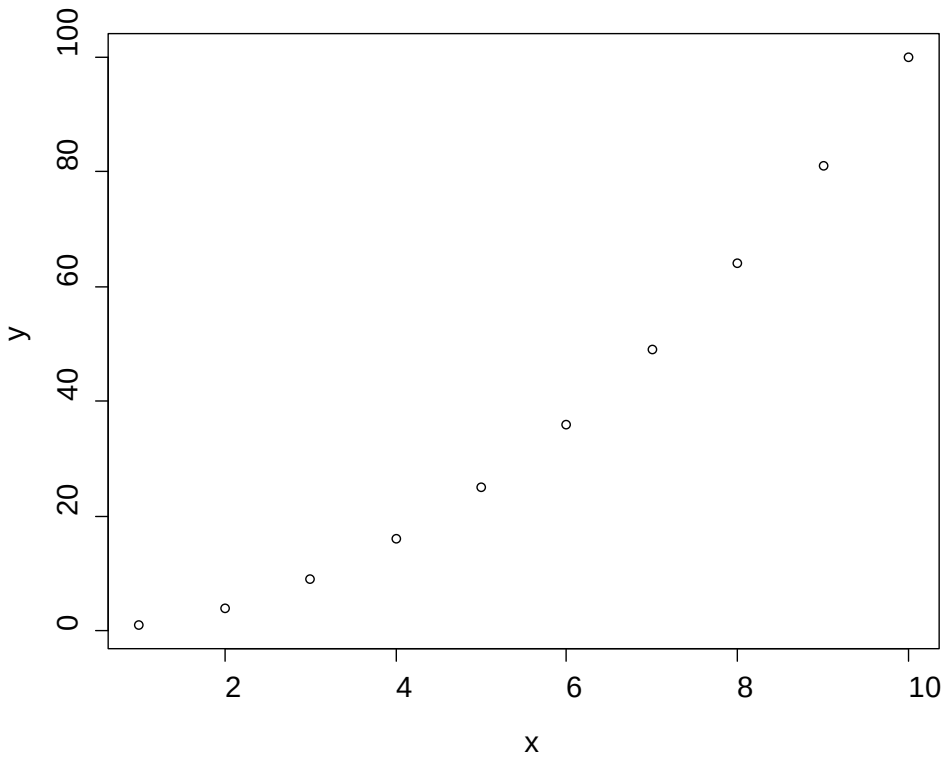
- 직선관계의 정도를 나타낸다.
- -1과 1사이의 값을 가진다.
- +: 같은 방향으로 움직이는 경향
- -: 반대 방향으로 움직이는 경향
- ±1에 가까울 수록 (즉, 절대값이 클수록) 강한 상관관계
- 0에 가까울 수록 관계없음
- ±1➔ 완벽한 직선관계를 의미

Kendal의 τ와 Spearman의 ρ

- 순위에 바탕을 둔 비모수적 방법
- 직선관계가 아니더라도 완벽한 상관관계가 있으면 1을 갖는다.

함수	내용
cor(matrix)	Pearson 상관계수 행렬
cor(matrix, method="kendall")	Kendall의 상관계수행렬
cor(matrix, method="spearman")	Spearman의 상관계수 행렬
cor(vector,vector)	상관계수
cor.test()	상관계수가 유의한지 검정

```
> x=1:10
> y=x^2
> plot(x,y)
>
> cor(x,y)
[1] 0.9745586
> cor(x,sqrt(y))
[1] 1
> cor(x,y,method="kendall")
[1] 1
> cor(x,y,method="spearman")
[1] 1
```



데이터분석

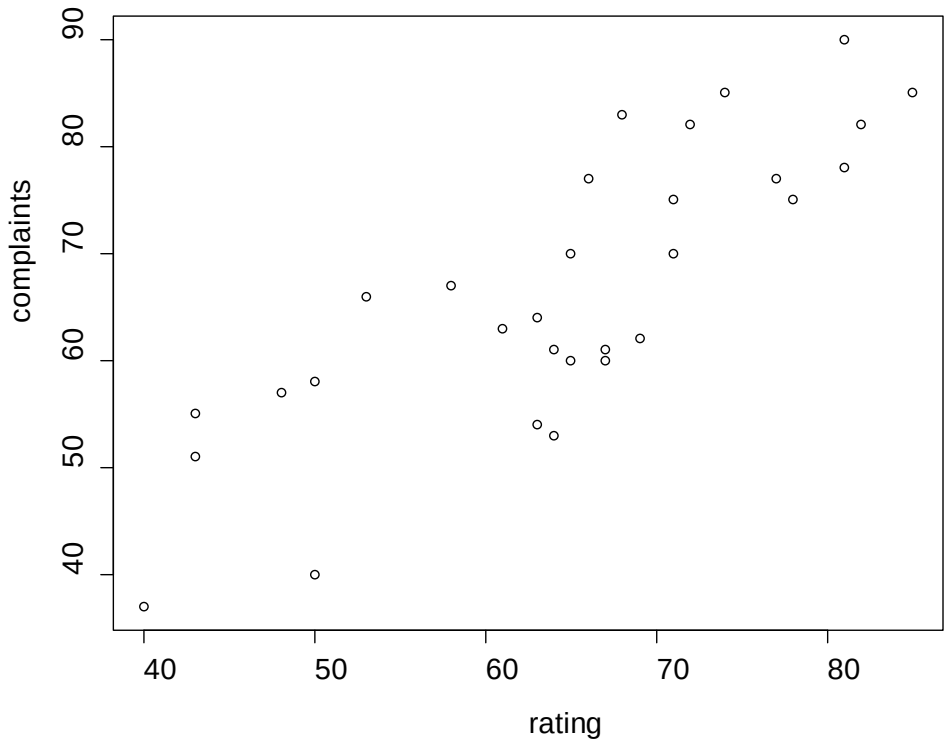
어느 큰 금융회사에서 30개 부서 에서 부서 당 약 35명의 직원으로부터의 설문결과를 부서별로 요약하였다. 데이터의 숫자는 해당 질문에 대해 긍정적으로 대답한 직원의 비율이다.

Y	rating	Numeric	Overall rating
X[1]	Complaints	Numeric	Handling of employee complaints
X[2]	Privileges	Numeric	Does not allow special privileges
X[3]	Learning	Numeric	Opportunity to learn
X[4]	Raises	Numeric	Raises based on performance
X[5]	Critical	Numeric	Too critical
X[6]	Advancel	Numeric	Advancement

```
> attitude
   rating complaints privileges learning raises critical advance
1      43         51         30       39     61       92      45
2      63         64         51       54     63       73      47
3      71         70         68       69     76       86      48
4      61         63         45       47     54       84      35
5      81         78         56       66     71       83      47
6      43         55         49       44     54       49      34
7      58         67         42       56     66       68      35
…
```

```
> cov(attitude)
             rating complaints privileges  learning    raises critical   advance
rating     148.17126  133.77931   63.46437  89.10460  74.68851 18.84253  19.42299
complaints 133.77931  177.28276   90.95172  93.25517  92.64138 24.73103  30.76552
privileges  63.46437   90.95172  149.70575  70.84598  56.67126 17.82529  43.21609
learning    89.10460   93.25517   70.84598 137.75747  78.13908 13.46782  64.19770
raises      74.68851   92.64138   56.67126  78.13908 108.10230 38.77356  61.42299
critical    18.84253   24.73103   17.82529  13.46782  38.77356 97.90920  28.84598
advance     19.42299   30.76552   43.21609  64.19770  61.42299 28.84598 105.85747
> cor(attitude)
              rating complaints privileges  learning    raises  critical   advance
rating     1.0000000  0.8254176  0.4261169 0.6236782 0.5901390 0.1564392 0.1550863
complaints 0.8254176  1.0000000  0.5582882 0.5967358 0.6691975 0.1877143 0.2245796
privileges 0.4261169  0.5582882  1.0000000 0.4933310 0.4454779 0.1472331 0.3432934
learning   0.6236782  0.5967358  0.4933310 1.0000000 0.6403144 0.1159652 0.5316198
raises     0.5901390  0.6691975  0.4454779 0.6403144 1.0000000 0.3768830 0.5741862
critical   0.1564392  0.1877143  0.1472331 0.1159652 0.3768830 1.0000000 0.2833432
advance    0.1550863  0.2245796  0.3432934 0.5316198 0.5741862 0.2833432 1.0000000
```

```
> with(attitude,cor.test(rating, complaints))

        Pearson's product-moment correlation

data:  rating and complaints
t = 7.737, df = 28, p-value = 1.988e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6620128 0.9139139
sample estimates:
      cor
0.8254176
```



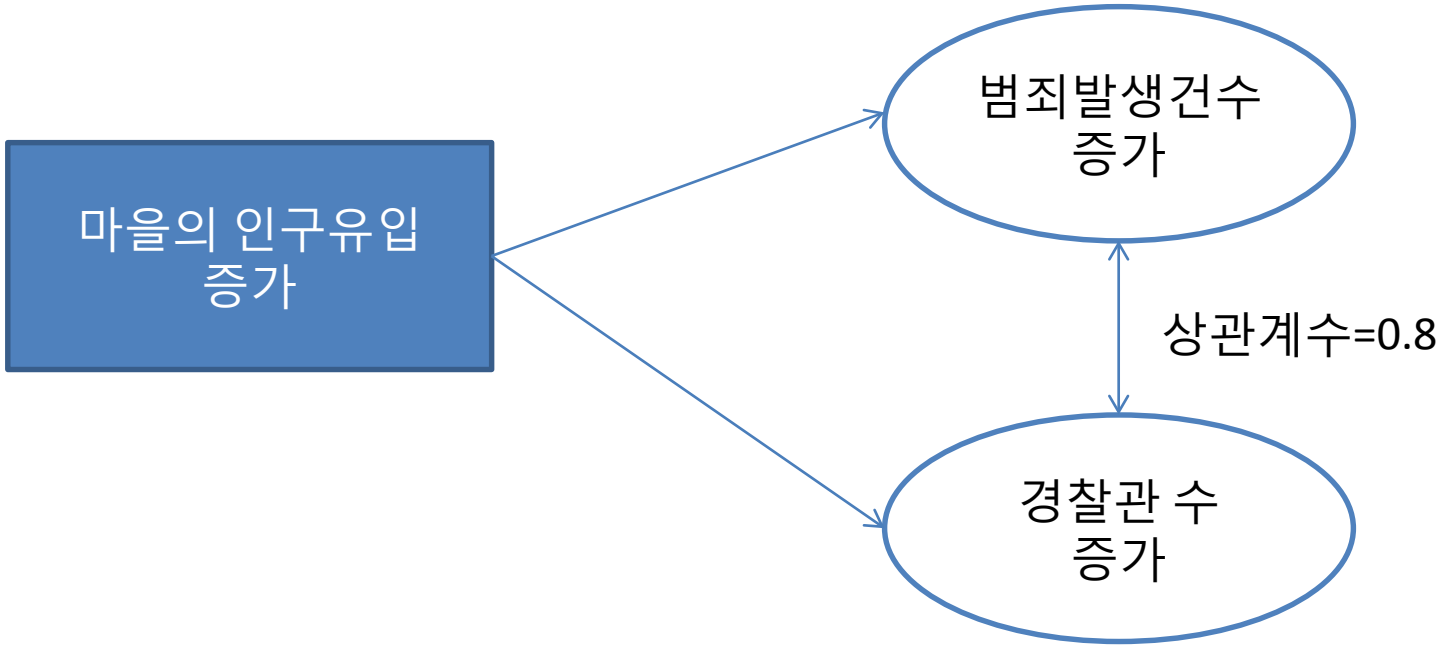
단순회귀분석

단순회귀분석 (Simple Linear Regression)

- 하나의 종속변수와 하나의 설명변수 간의 관계를 직선으로 표현하는 방법
- 종속변수: 예측될 변수
- 설명변수 (독립변수): 종속변수를 예측하는데 활용될 변수

상관분석 vs 회귀분석

- 상관분석
  - 두 변수 간의 선형관계의 강도 측정
  - 인과관계 없음
  - False 상관관계 유의



- 회귀분석
  - 원인이 되는 변수 (설명변수)에 따른 종속변수의 결과 예측 (의존적 관계)
  - 둘 이상의 변수들 간의 관계
  - 상관관계 포함
  - 인과관계는 통계학의 범주를 넘어서서 이론적인 선험적인 고려가 선행되어야 한다.

모형

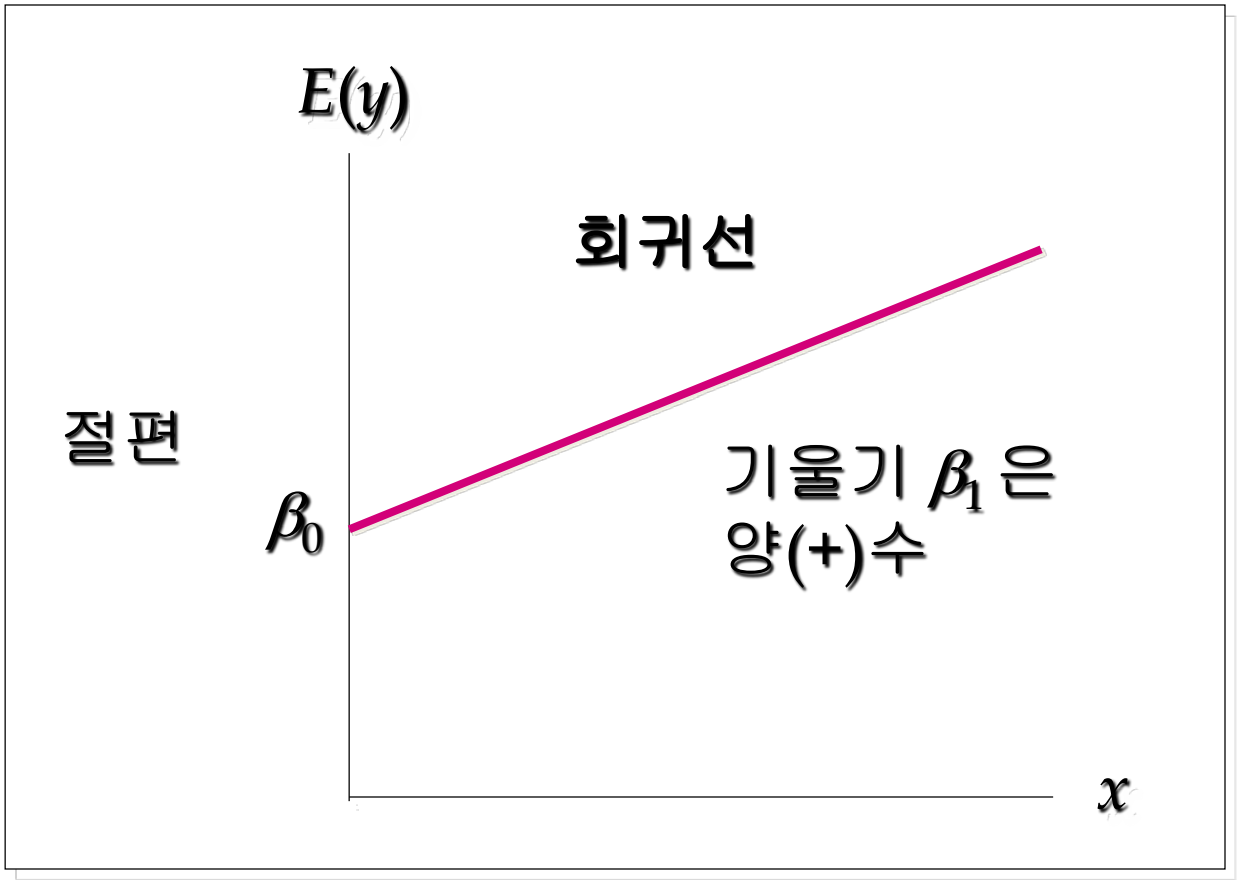
$$y = \beta_0 + \beta_1 x + \varepsilon$$

여기서:

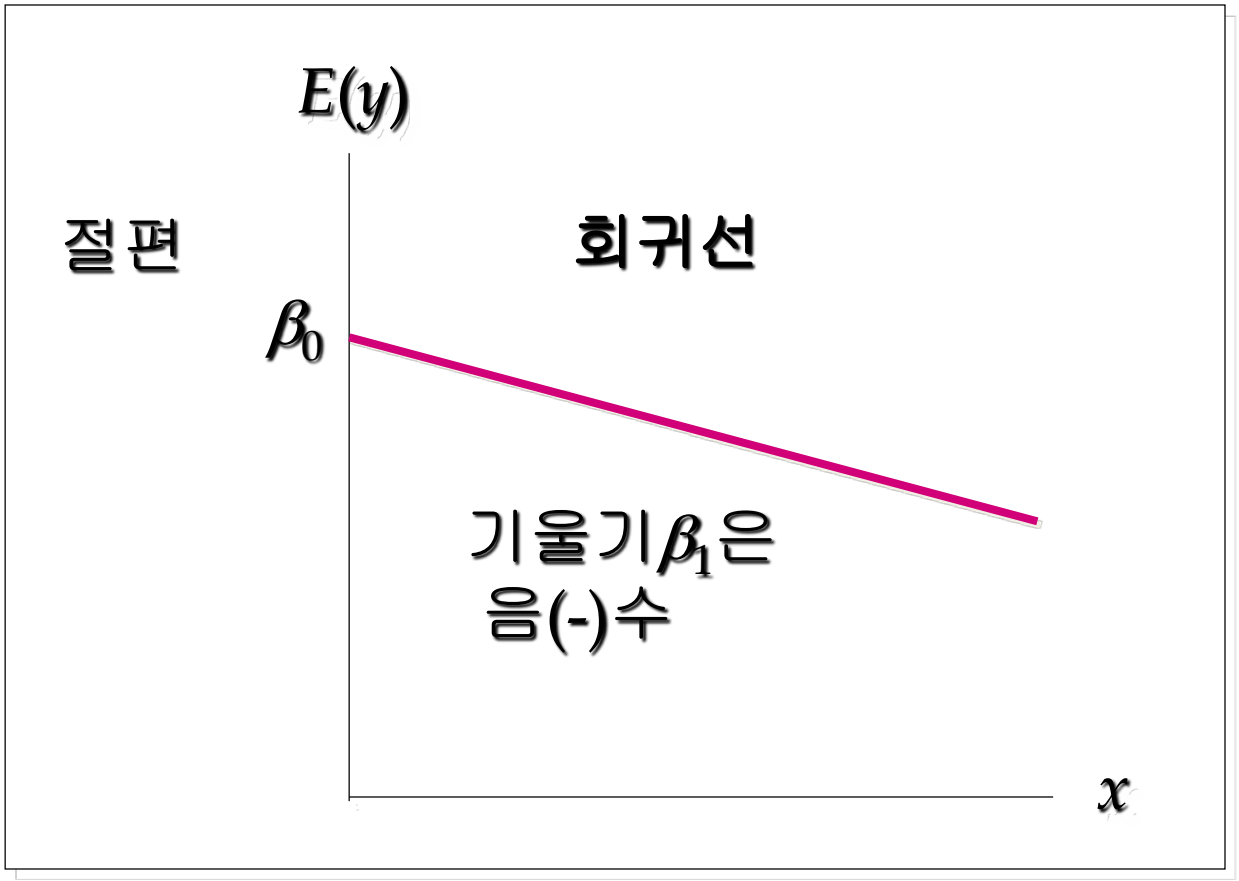
β_0와 β_1은 모형의 모수이고,

ε(그리스문자 epsilon)은 오차항(error term)이라 불리는 확률변수이다.

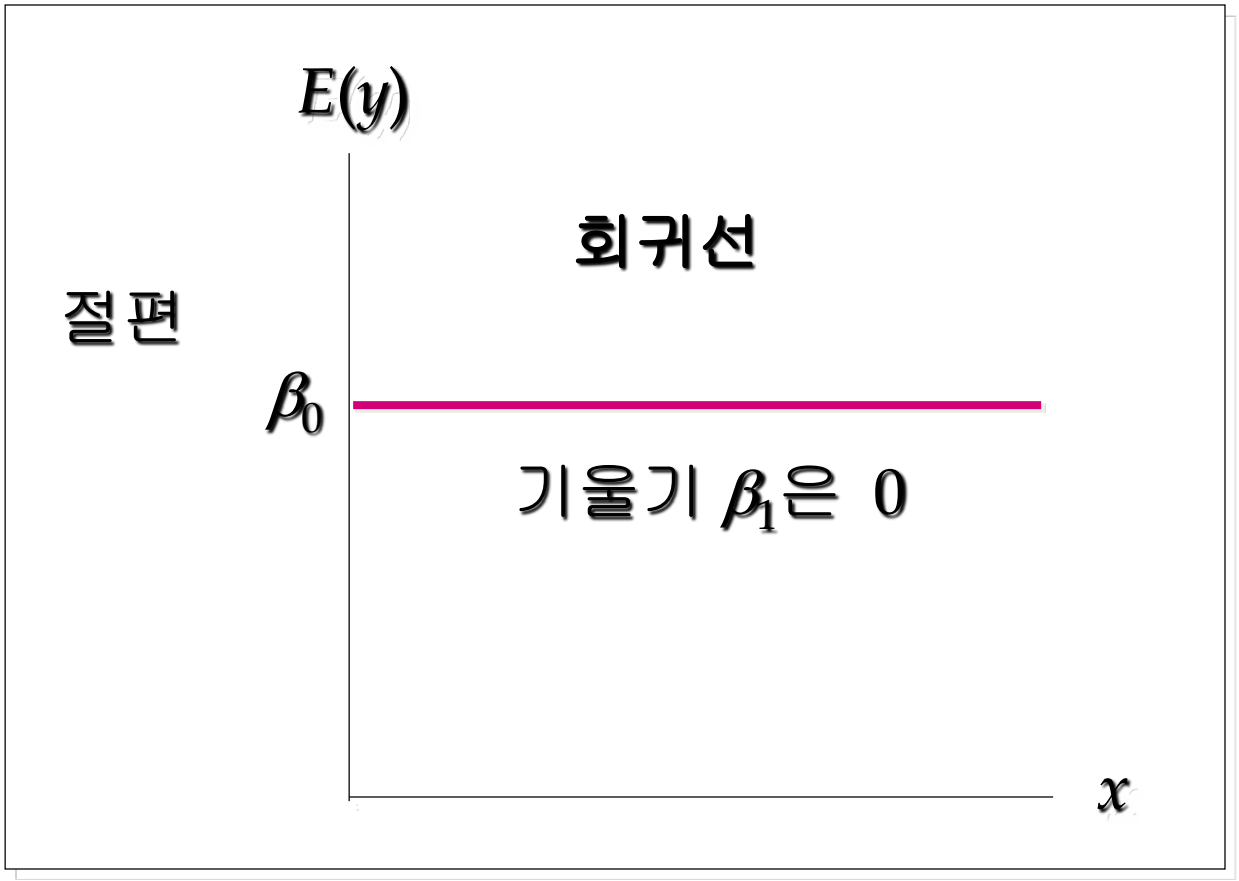
- 양(+)의 선형관계



- 음(-)의 선형 관계



- 관계 없음



추정단순선형회귀식

- 추정(된)단순선형회귀식:

$$\hat{y} \equiv b_0 + b_1 x$$

  - 그래프는 추정회귀선이라 한다.
  - b_0 은 y절편이다.
  - b_1 은 기울기이다.
  - $\hat{y}$ 주어진 x값에 대한 y의 추정값이다.

최소자승법(least squares method)

- 최소자승기준

$$\min \sum (y_i - \hat{y}_i)^2$$

여기서 :

y_i = i 번째 관찰값에 대한 종속변수의 관찰값

$\hat{y}_i$ = i 번째 관찰값에 대한 종속변수의 추정값

최소자승법

추정회귀식의 기울기

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

추정회귀식의 *y*절편

$$b_0 \equiv \bar{y} - b_1 \bar{x}$$

여기서:

x_i = i 번째 관찰값에 대한 독립변수의 값

y_i = i 번째 관찰값에 대한 종속변수의 값

$\bar{x}$ = 독립변수의 평균값

$\bar{y}$ = 종속변수의 평균값

n = 관찰점 수

예) CARS

- 차의 속도와 급브레이크를 밟았을 때 멈추기까지 걸린 거리

```
> cars=read.csv("cars.csv")
> cars
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
…
```

회귀분석 in R

R 명령어	내용
lm(종속변수 ~설명변수, data)	설명변수를 종속변수에 회귀분석
plot(lm())	회귀분석관련 그래프 출력
summary(lm())	회귀분석 결과 summary
abline(intercept, slope) abline(lm())	기존의 그래프에 직선 추가. Intercept과 slope를 인수로 넣거나 lm결과 를 인수로 넣을 수 있음

회귀분석 in R

```
> out=lm(dist~speed,data=cars)
> summary(out)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511,     Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

$b_1 =$

$b_0 =$

회귀계수에 대한 검정

$$Test\ H_0: \beta_1 = 0\ \ vs\ \ H_a: \beta_1 \neq 0$$

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

결과 해석

b_1: 속력이 1만큼 증가했을 때 거리는 _____만큼 증가한다.
b_0: 속력이 0일 때 거리는 _____이다?

- 귀무가설 H_0: β_1=0을 기각하여 x와 y의 관계가 유의하다고 하더라도 x와 y 간에 원인-결과 관계가 존재한다고 결론 내릴 수는 없다.
- H_0: β_1=0을 기각하고 통계적 유의성만 검정할 수 있기 때문에 x와 y의 관계가 선형이라고 결론내릴 수 없다.
- Y절편(b_0)에 대한 해석은 설명변수 자료의 범위가 0을 포함할 때만 의미가 있다.

결정계수 (R^2)

- 회귀모형의 설명력을 평가
- 언제나 0과 1사이
- 0: 회귀모형이 종속변수의 변동량을 전혀 설명하지 못한다
- 1: 회귀모형이 종속변수의 변동량을 100% 설명한다.
- 단순회귀분석에서는 두 변수 사이의 상관계수의 제곱과 일치한다.

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511,     Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

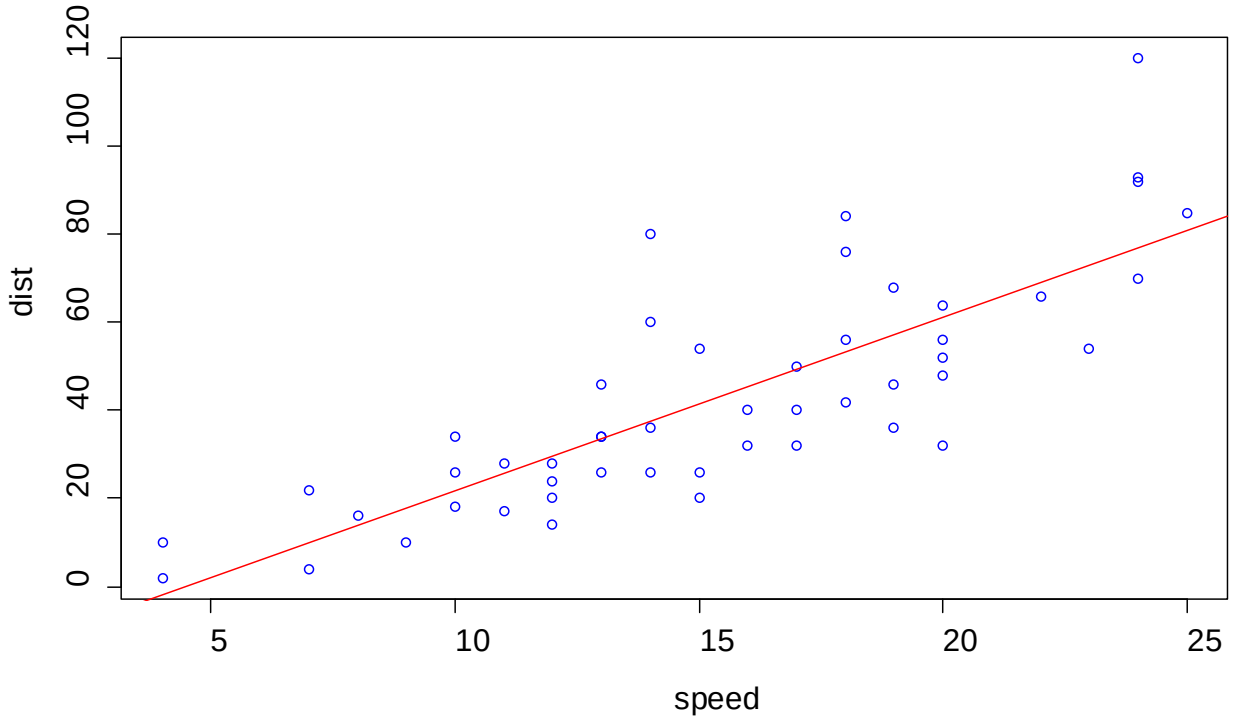
유의성검정: F-test

- 단순회귀분석에서는 $H_0: \beta_1 = 0 \ vs \ H_a: \beta_1 \neq 0$를 검정하는 t-test 와 동일

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared: 0.6511,     Adjusted R-squared: 0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

산점도와 회귀선

```
> plot(dist~speed,data=cars,col="blue")
> abline(out,col="red")
```



No Intercept Model

- 속도가 0이면 멈추기 까지 걸린 거리도 0인 것이 당연하다.➔ β_0 을 0으로 고정하자

```
> summary(lm(dist~speed+0,data=cars))

Call:
lm(formula = dist ~ speed + 0, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-26.183 -12.637  -5.455   4.590  50.181

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
speed   2.9091     0.1414   20.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.26 on 49 degrees of freedom
Multiple R-squared: 0.8963, Adjusted R-squared: 0.8942
F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16
```