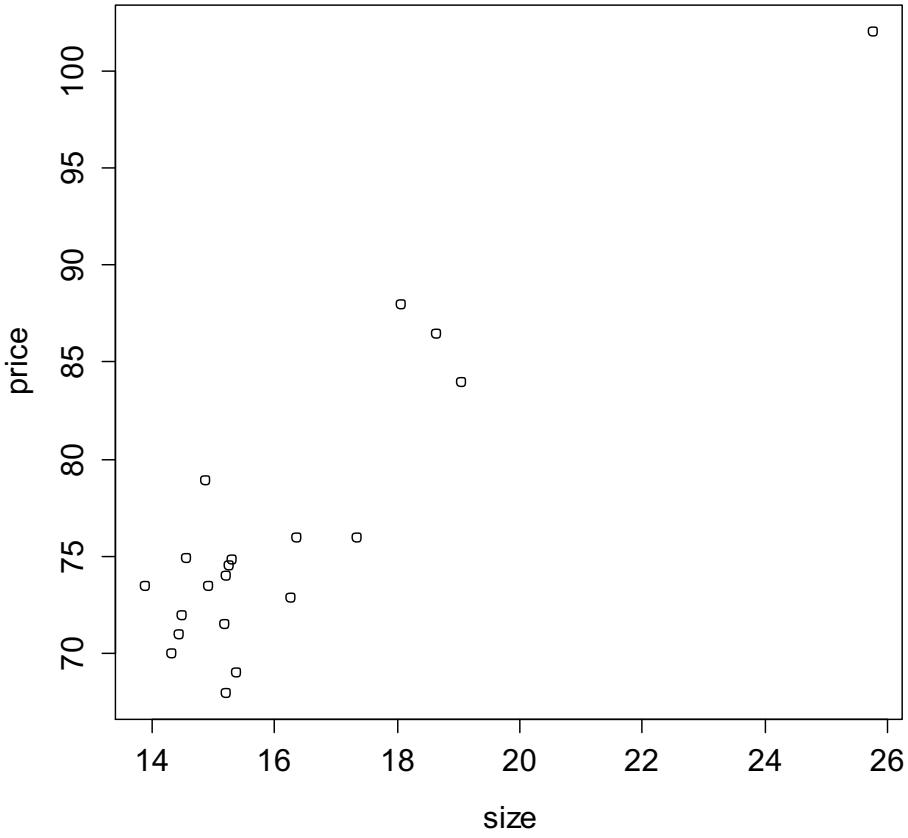
상관분석

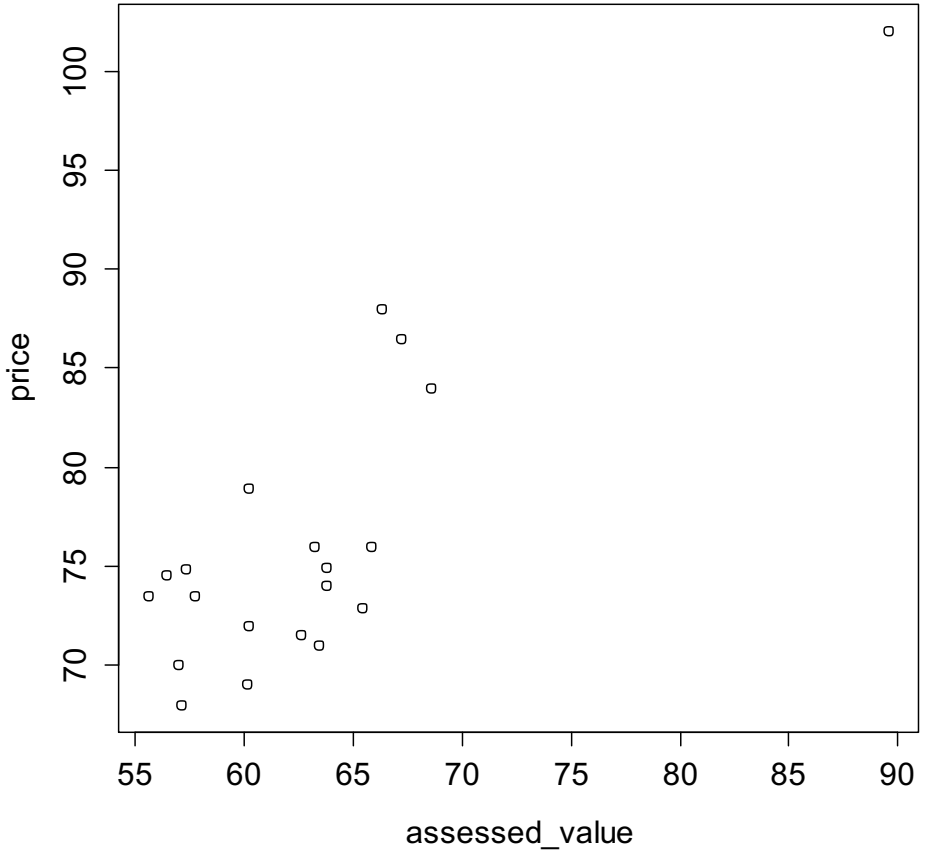
예 : 주택 가격

- Milwaukee, Wisconsin의 20개 주택 가격에 대한 자료이다.

Total Dwelling Size ($100\ ft^2$)	Assessed Value ($1000)	Selling Price ($1000)
15.31	57.3	74.8
15.2	63.8	74
16.25	65.4	72.9
14.33	57	70
14.57	63.8	74.9
17.33	63.2	76
14.48	60.2	72
14.91	57.7	73.5
15.25	56.4	74.5
13.89	55.6	73.5
15.18	62.6	71.5
14.44	63.4	71
14.87	60.2	78.9
18.63	67.2	86.5
15.2	57.1	68
25.76	89.6	102
19.05	68.6	84
15.37	60.1	69
18.06	66.3	88
16.35	65.8	76

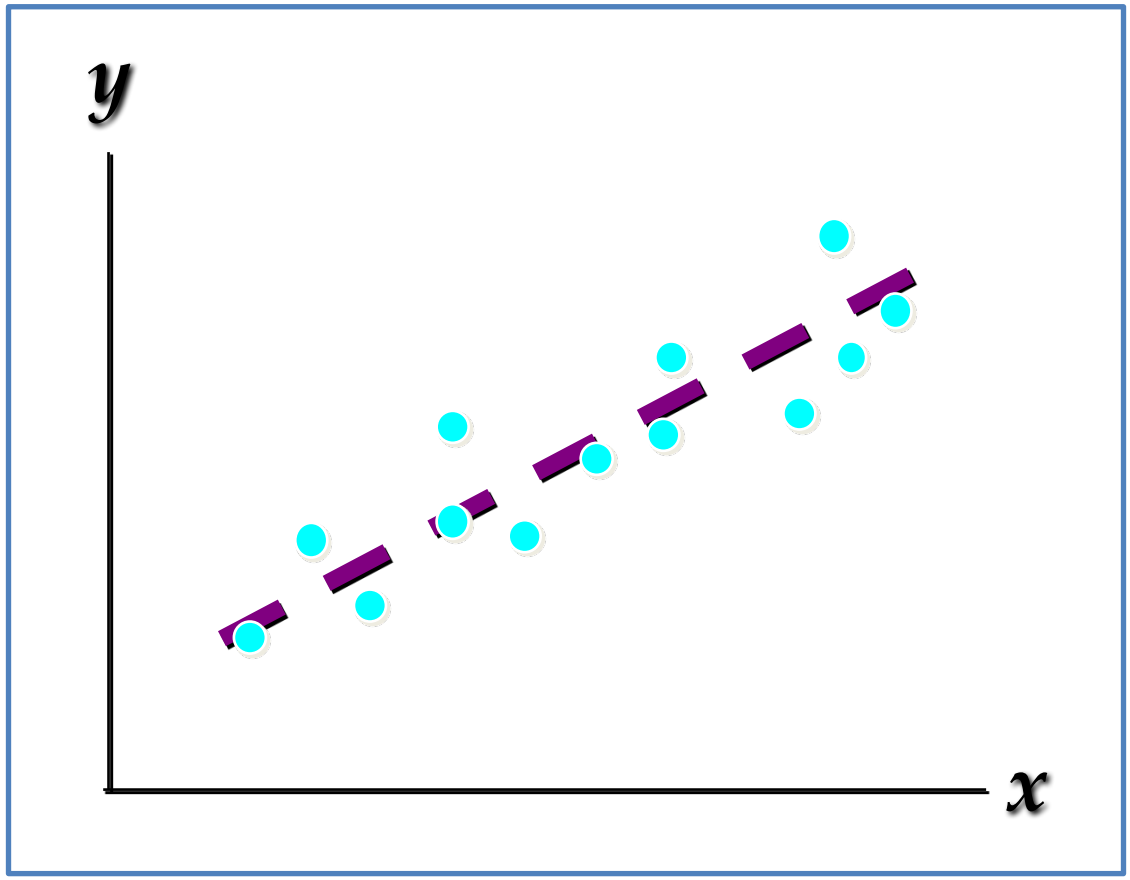
산점도


price
70
75
80
85
90
95
100
14
16
18
20
22
24
26
size


price
70
75
80
85
90
95
100
55
60
65
70
75
80
85
90
assessed_value

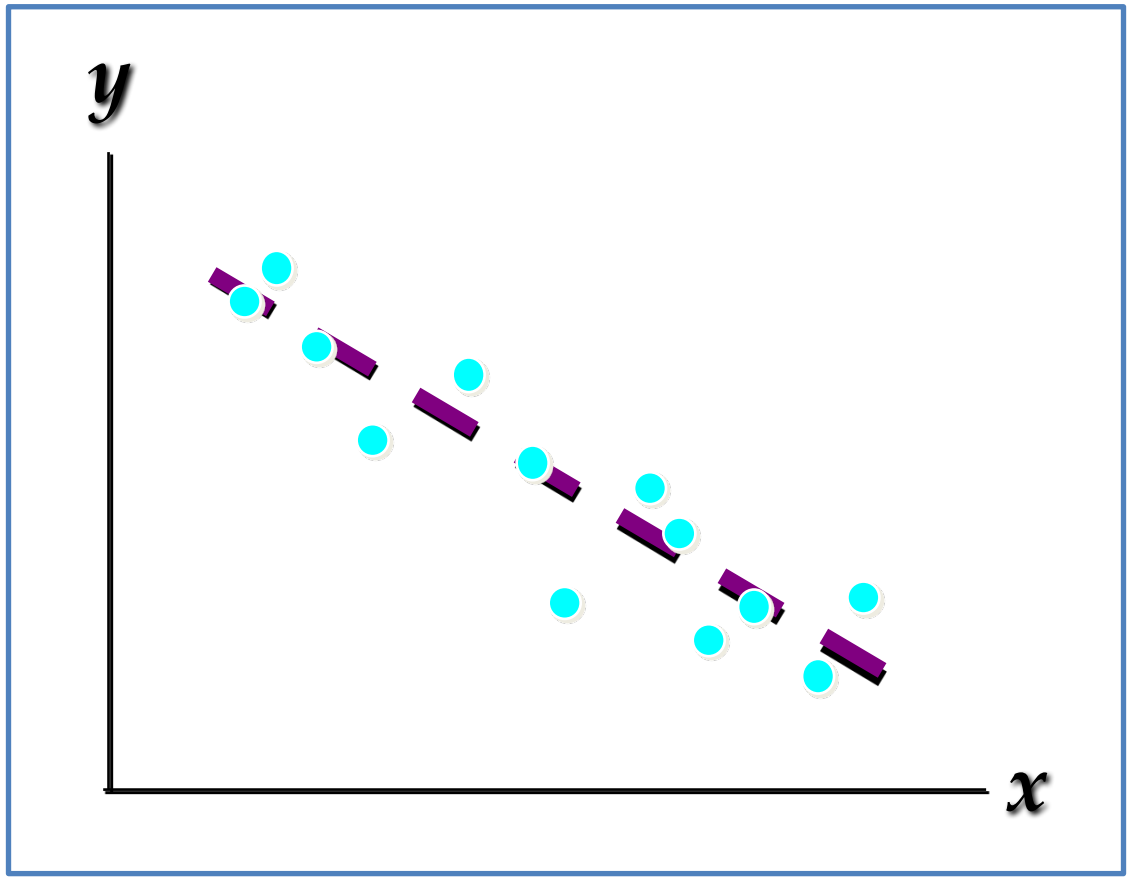
산점도

- 정의 관계



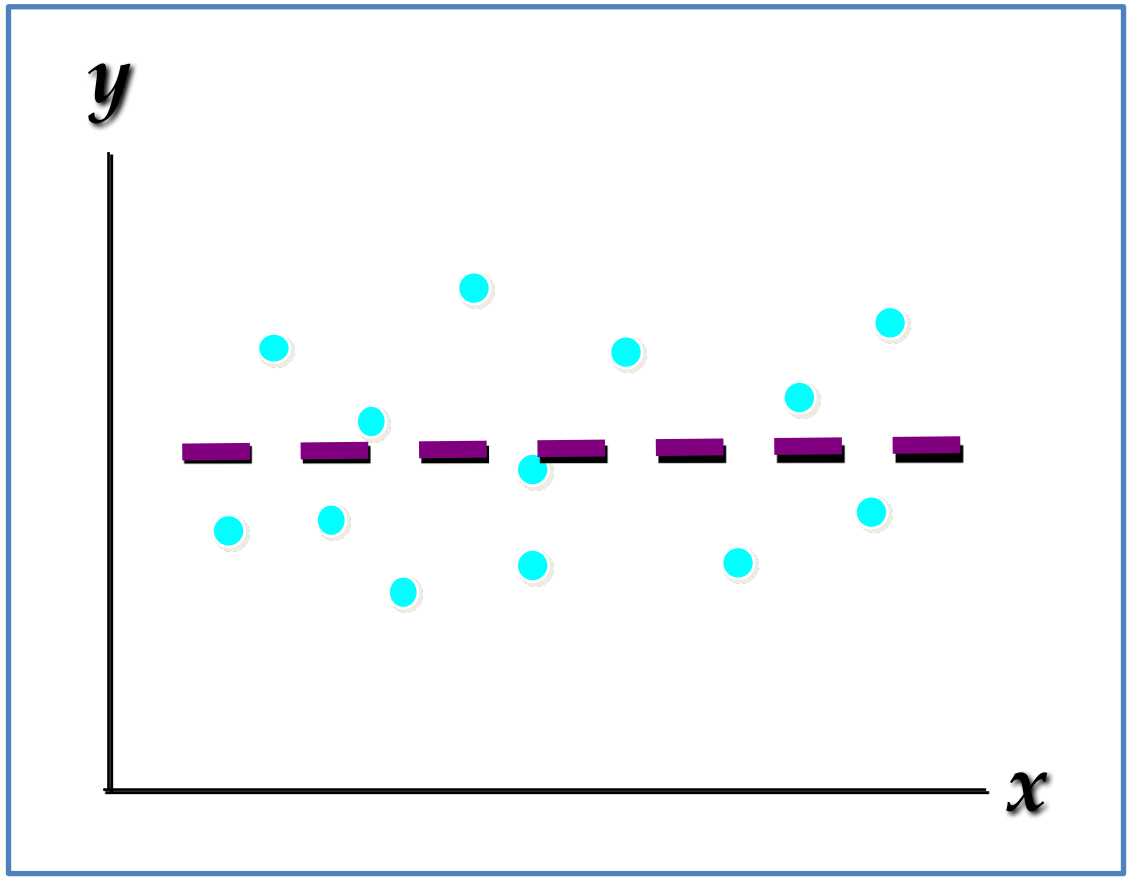
산점도

- 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

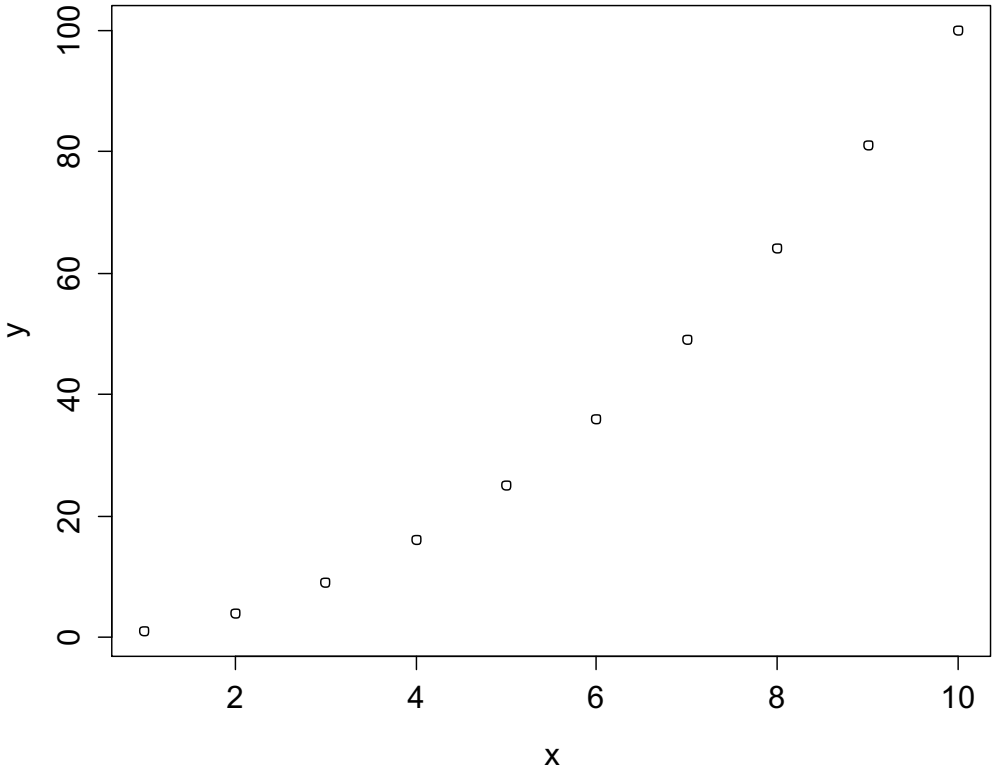
- 직선관계의 정도를 나타낸다.
- -1과 1사이의 값을 가진다.
  - 양수: 같은 방향으로 움직이는 경향
  - 음수: 반대 방향으로 움직이는 경향
- ± 1에 가까울 수록 (즉, 절대값이 클수록) 강한 상관관계
- 0에 가까울 수록 관계없음
- ± 1➔ 완벽한 직선관계를 의미

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
```



예: 직원 설문조사

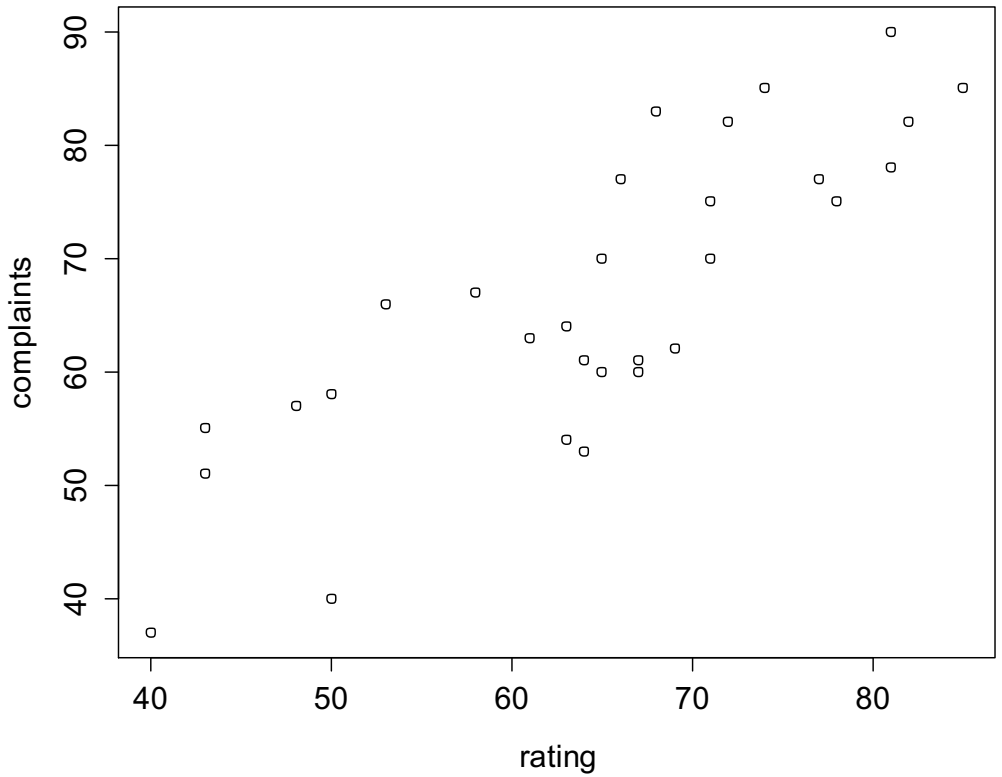
어느 금융회사에서 30개 부서 에서 부서 당 약 35명의 직원으로부터의 설문결과를 부서별로 요약하였다. 데이터의 숫자는 해당 질문에 대해 긍정적으로 대답한 직원의 비율이다.

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
```

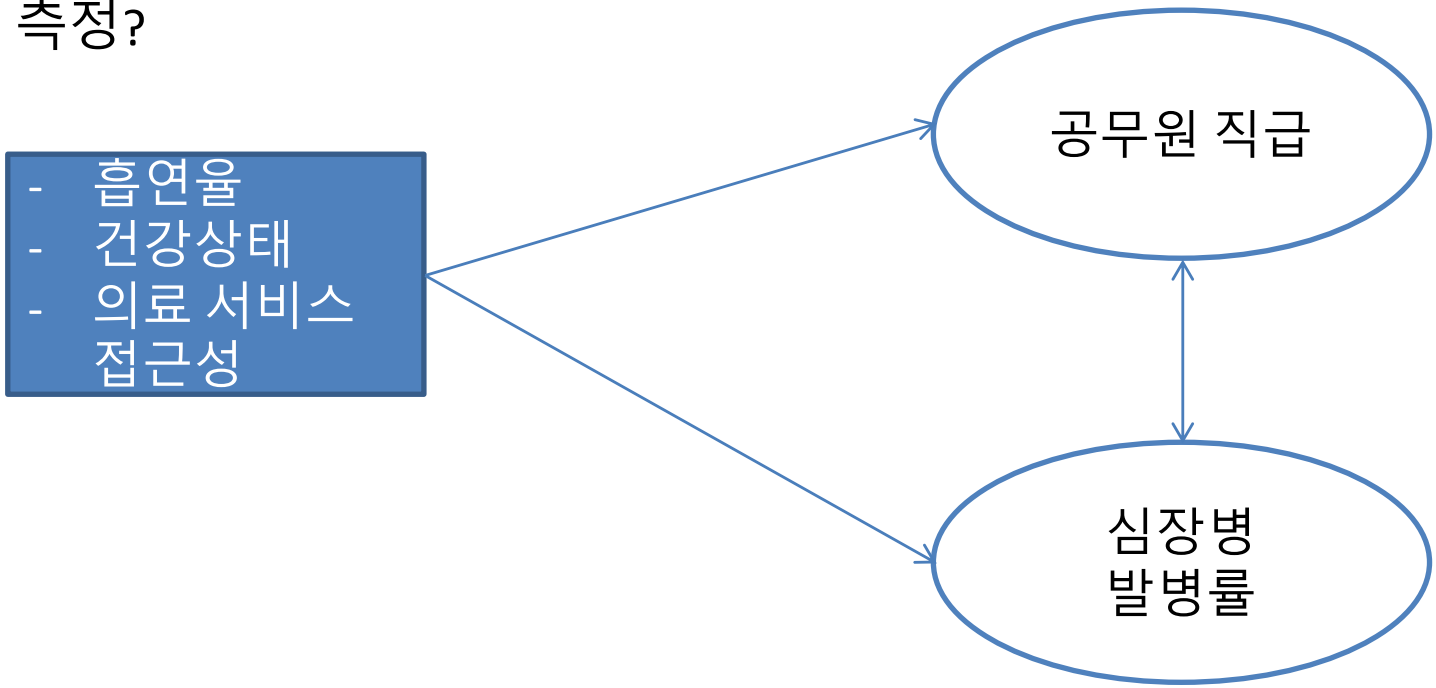
```
> with(attitude,cor.test(rating, complaints))

        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
```



상관관계 vs. 인과관계

- 업무 스트레스로 인한 사망?
  - CEO보다 비서의 사망률이 더 높음
  - 하급 공무원의 심장병 발병률이 고급 공무원의 발병률보다 높음
  - 자신이 맡은 직무에 결정권이 적을 수록 업무 스트레스를 받음

- 무작위 실험이 불가능
  - 서로 다른 직무에 무작위로 배정하고 수년간 그 일을 시킨 다음 사망률 측정?



단순회귀분석

단순회귀분석 (Simple Linear Regression)

- 하나의 종속변수와 하나의 설명변수 간의 관계를 직선으로 표현하는 방법
- 원인이 되는 변수에 따른 종속변수의 결과 예측 (의존적 관계)
  - 종속변수: 예측될 변수
  - 설명변수 (독립변수): 종속변수를 예측하는데 활용될 변수
- 인과관계는 통계학의 범주를 넘어서서 이론적인 선험적인 고려가 선행되어야 함

단순회귀모형

$$\boldsymbol{y = \beta_0 + \beta_1 x + \epsilon}$$

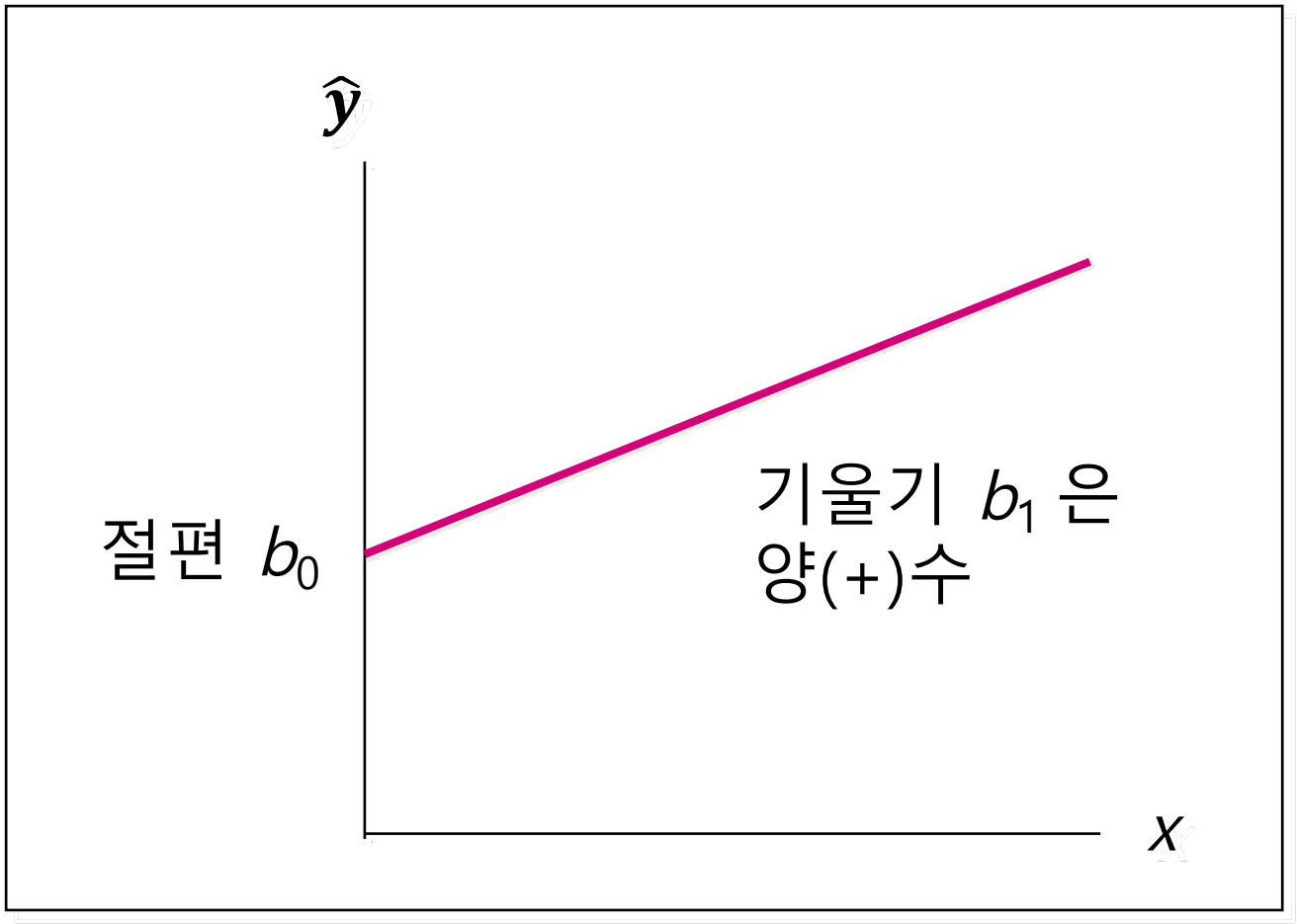
β_0: y-절편 (모수)

β_1: 기울기 (모수)

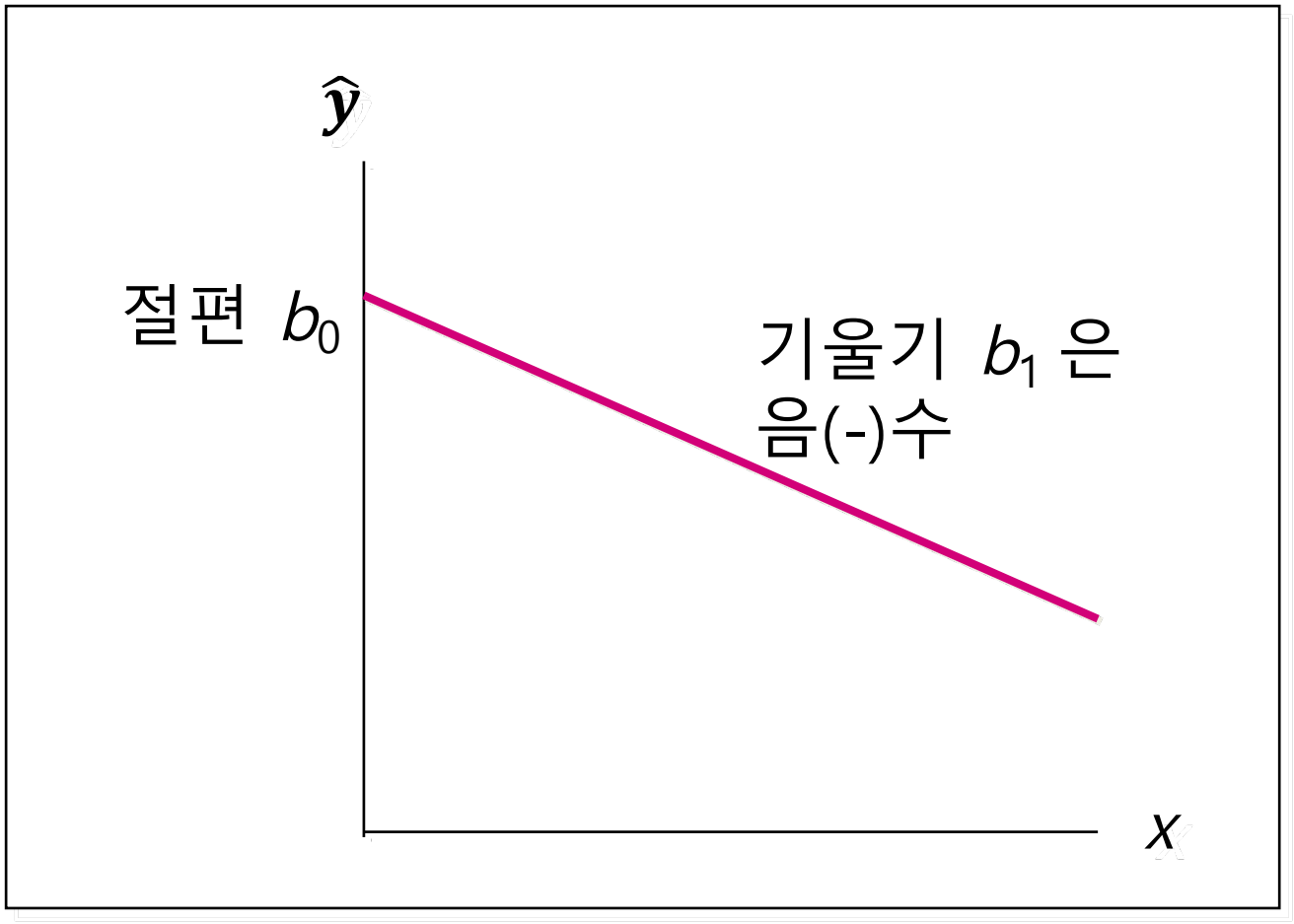
ε: 오차항 (확률변수: 평균 0, 분산 σ^2)

추정된 회귀모형: $\boldsymbol{\hat{y} = b_0 + b_1 x}$

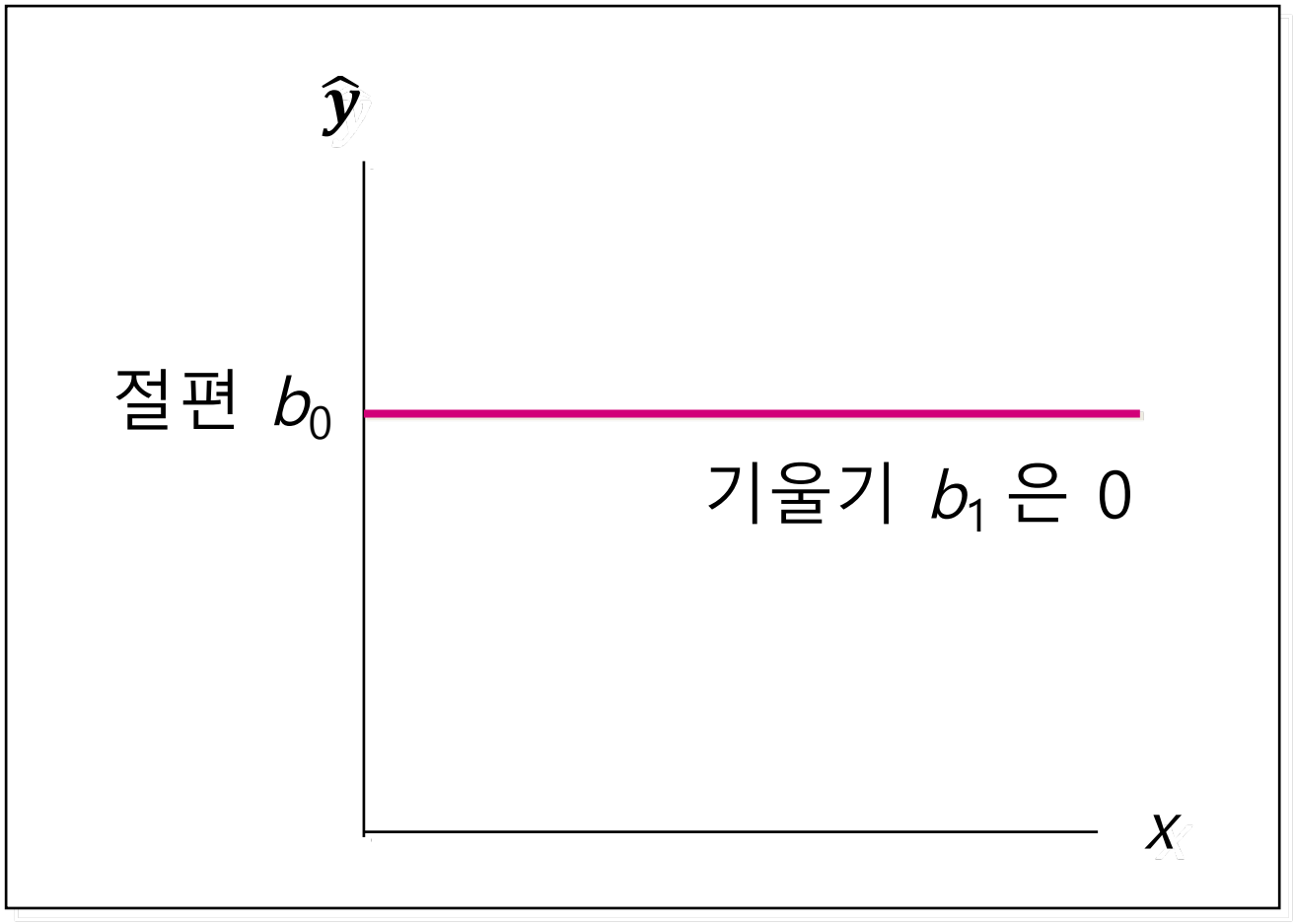
- 양(+)의 선형관계



- 음(-)의 선형관계



- 관계없음



회귀모형의 가정

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

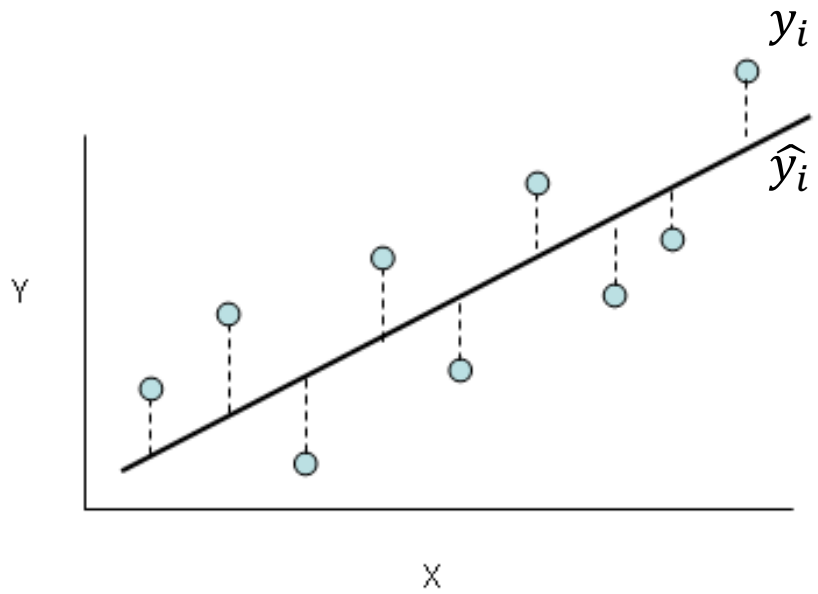
- y 값들은 서로 독립이다.
- y 와 x는 선형관계이다.
- $y|x$ (같은 x 값에 대한 y)의 분산은 동일하다.
- $y|x$는 정규분포를 따른다.

회귀식을 어떻게 추정하나?
최소자승법(least squares method)

$$\min \sum_i (y_i - \hat{y_i})^2$$

y_i = i 번째 관찰값에 대한 종속변수의 관찰값

$\hat{y_i}$ = i 번째 관찰값에 대한 종속변수의 추정값



최소자승법

추정회귀식의 기울기

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

추정회귀식의 *y*절편

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

여기서:

x_i = i 번째 관찰값에 대한 독립변수의 값

y_i = i 번째 관찰값에 대한 종속변수의 값

$\bar{x}$ = 독립변수의 평균

$\bar{y}$ = 종속변수의 평균

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

회귀계수 추정과 해석

```
> 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
```

$b_1 =$

$b_0 =$

b_1: 속력이 1만큼 증가했을 때 거리는 _____ 만큼 증가한다.

b_0: 속력이 0일 때 거리는 _____ 이다?

회귀계수에 대한 검정: t-test

- 설명변수가 종속변수에 대해 유의한 설명력을 가지는지 검정
- $H_0: \beta_1 = 0 \ \ vs \ \ H_a: \beta_1 \neq 0$

```
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 해석의 주의점

- 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.

모형의 유의성검정: F-test

- 회귀 모형이 종속변수의 변동을 설명하는 데 유의한지 검정
- H_0: 회귀모형이 유의하지 않다.
- H_1: 회귀모형이 유의하다.
- 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
```

ANOVA Table

Source	Df	Sum of Squares (SS)	Mean Sqaure (MS)	Variance Ratio (F)
Regression	1	SSR	MSR	MSR/MSE
Residual	N-2	SSE	MSE	
Total	N-1	SST		

SSR=$\sum_i(\hat{y_i}-\bar{y})^2$: 회귀식에 의해 설명되는 변동량
SSE=$\sum_i(y_i-\hat{y_i})^2$: 회귀식에 의해 설명되지 않는 변동량
SST=$\sum_i(y_i-\bar{y_i})^2$: 총 변동량

SST=SSR+SSE

```
> anova(out)
Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value   Pr(>F)
speed      1  21186 21185.5  89.567 1.49e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

결정계수 (R^2)

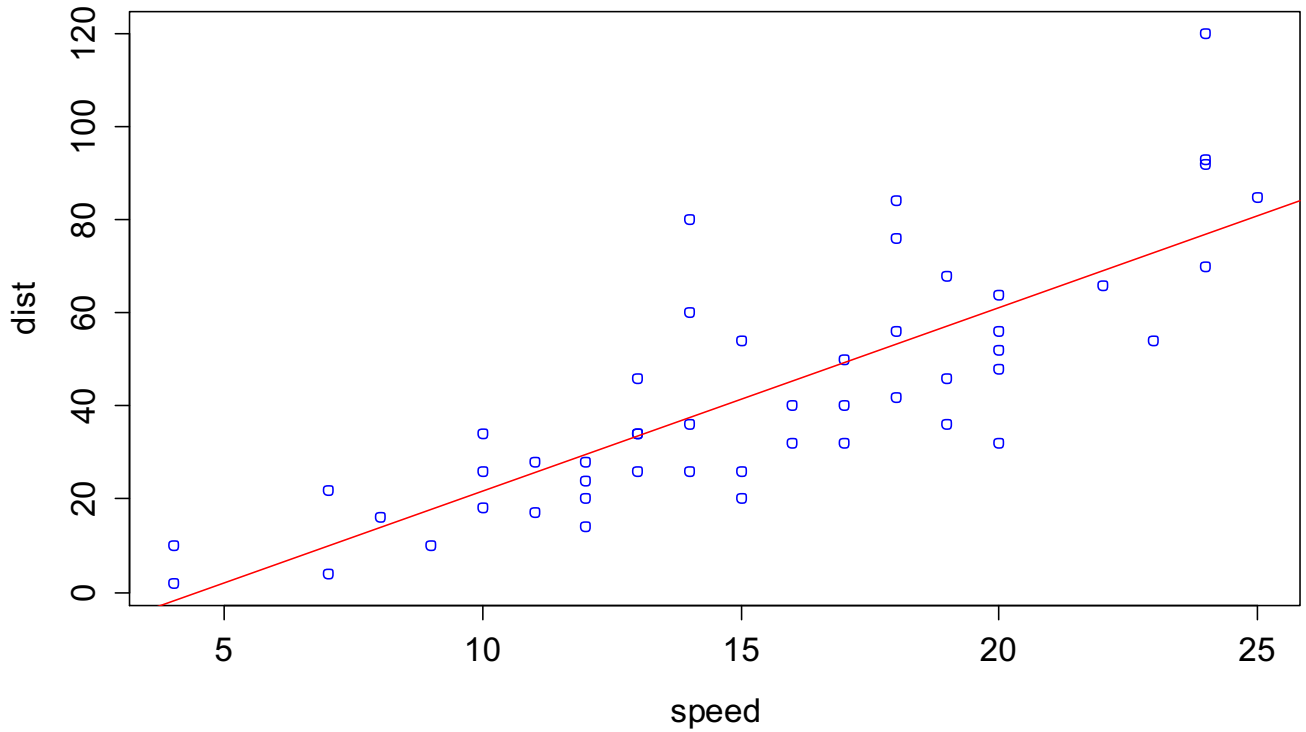
$$R^2 = \frac{SSR}{SST}$$

- 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
```

산점도와 회귀선

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



단순회귀분석

회귀진단

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
```

회귀진단

- 만약 회귀모형이 제대로 설정되고 추정되었다면

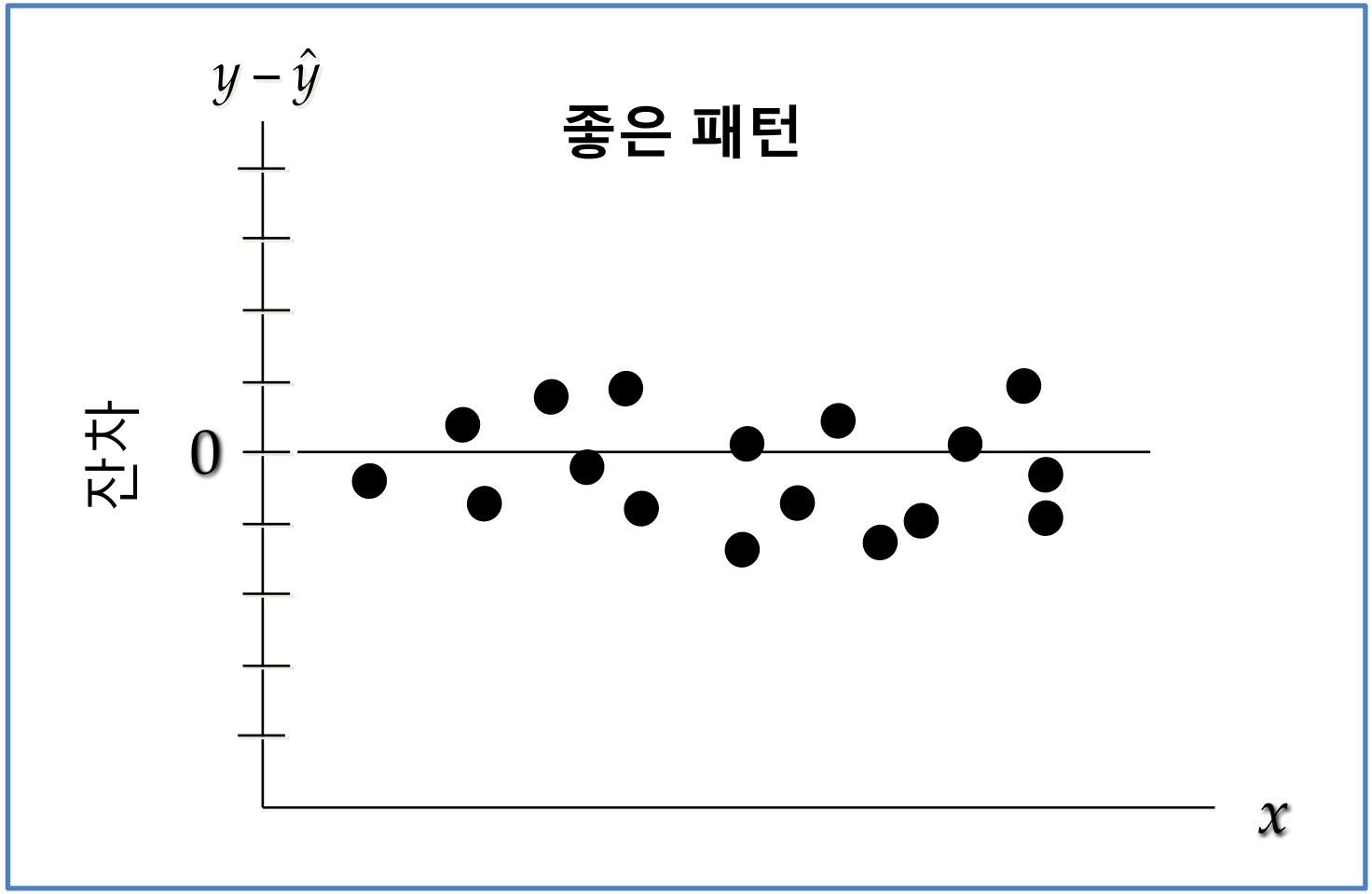
오차항 ϵ= 찌꺼기

오차항이 특정한 패턴을 보인다면 무언가 중요한 정보가 모형에 포함되지 않았다는 의미

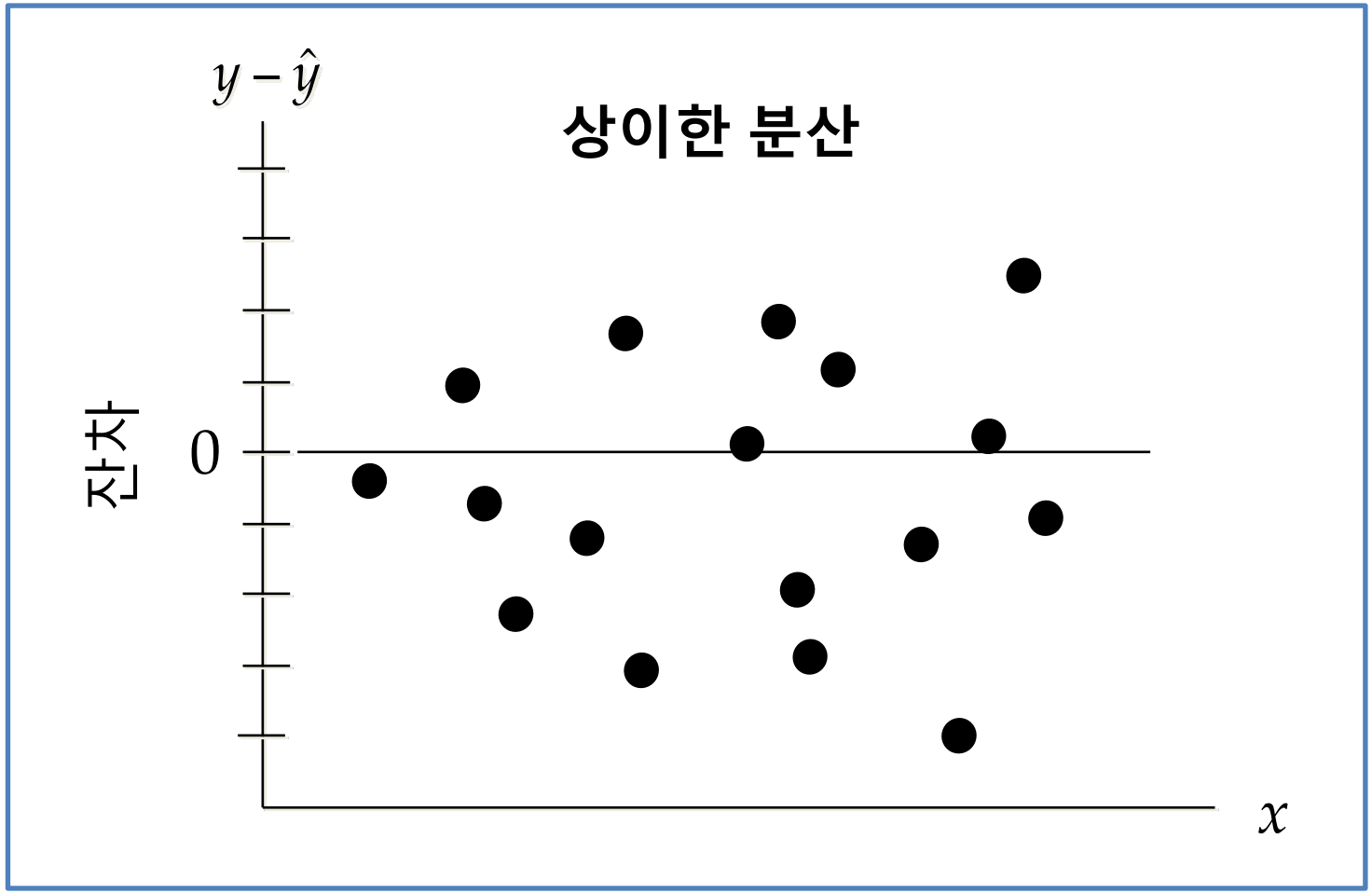
오차항에 대한 가정

1. 오차항 ε 은 평균이 '0'인 확률변수이다.
2. ε의 분산은 모든 x값에 대해 동일하다.
3. ε 값들은 서로 독립적이다.
4. 오차항 ε 은 정규분포를 이루는 확률변수이다.

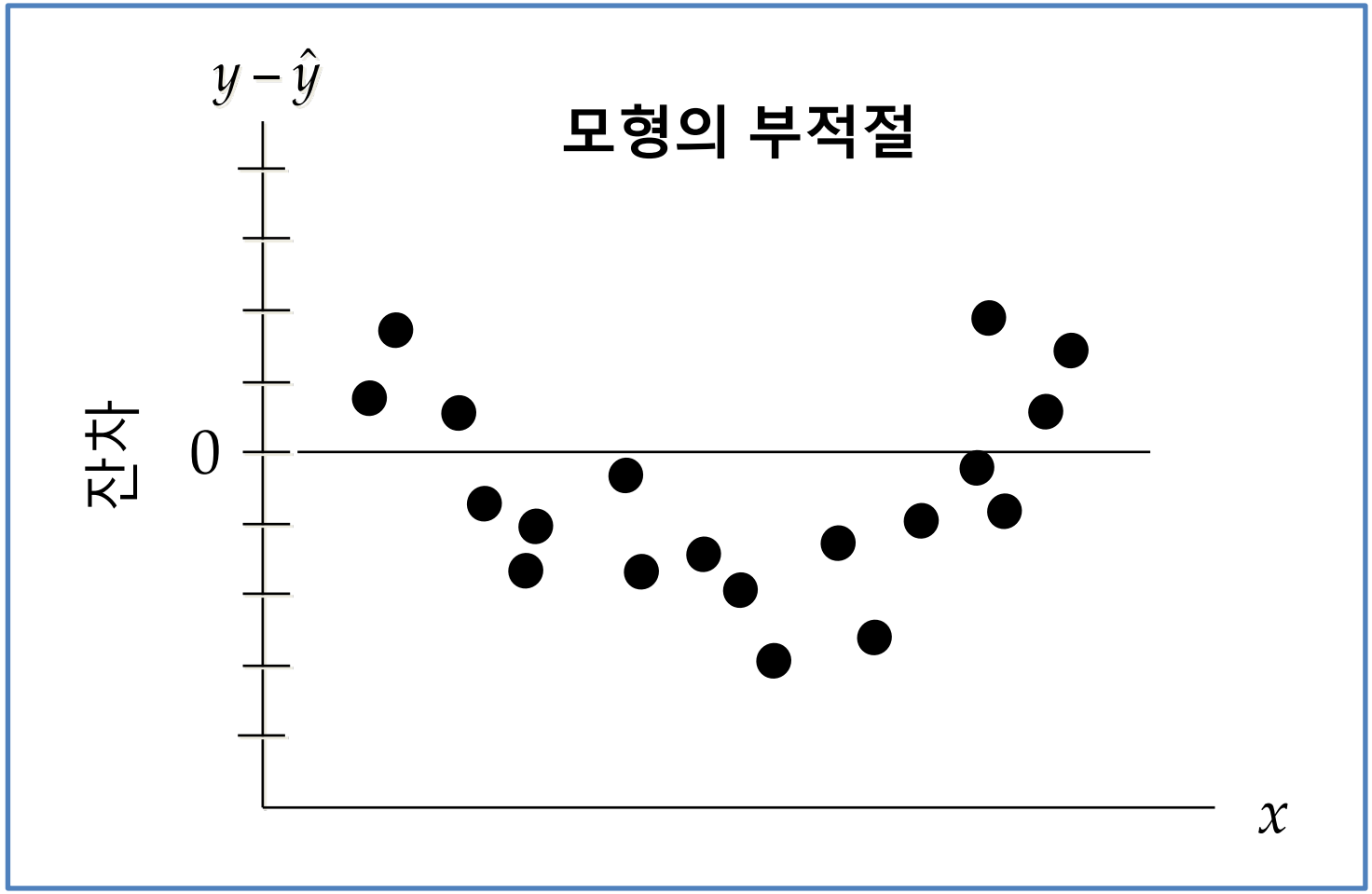
잔차도 (residual plot)



잔차도 (residual plot)

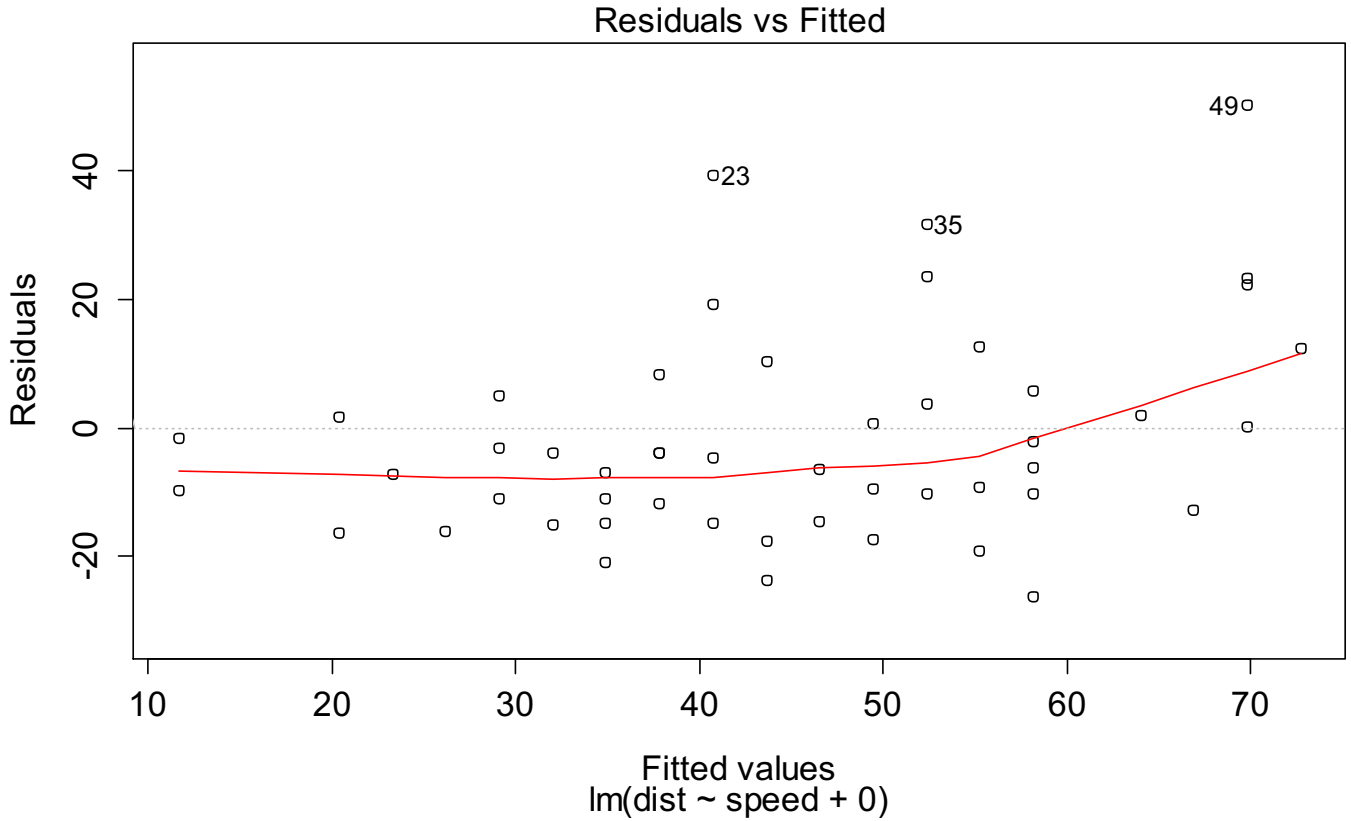


잔차도 (residual plot)



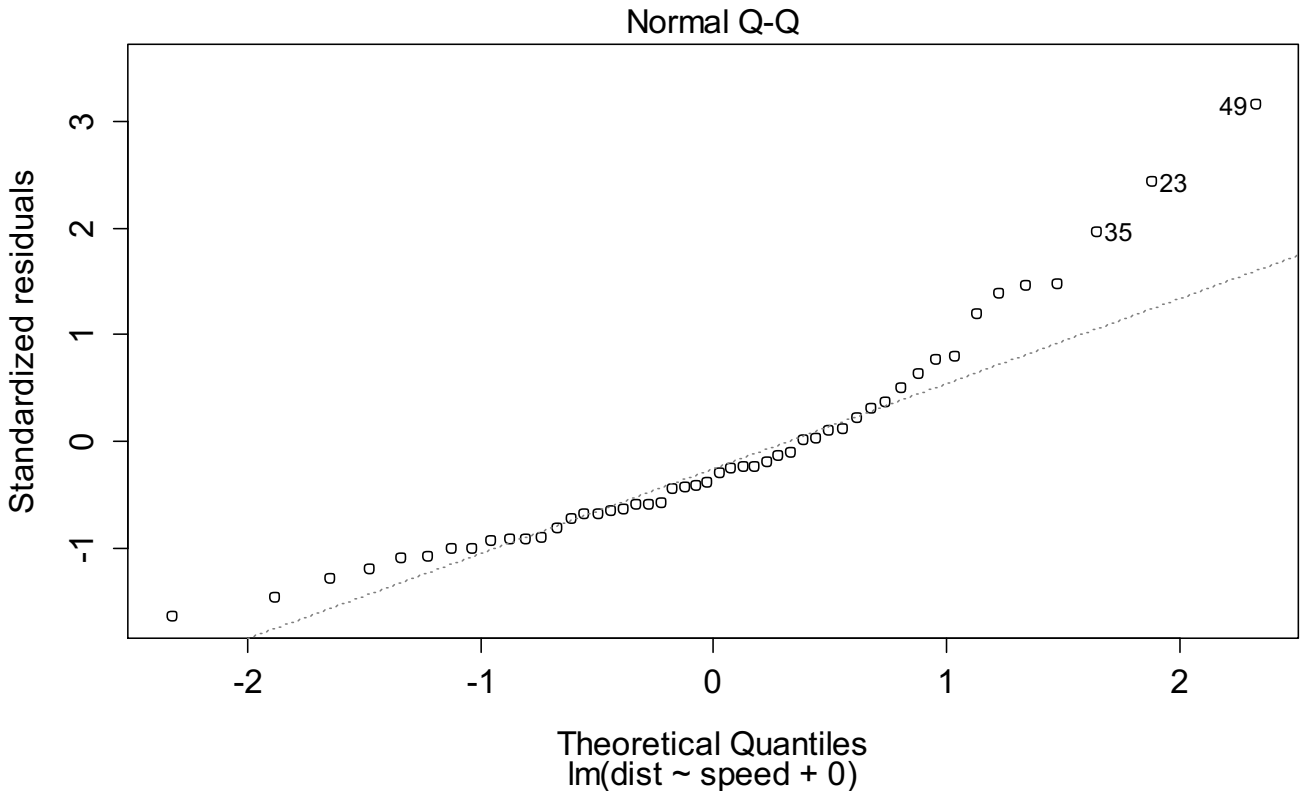
잔차도: CARS, No intercept 모형

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

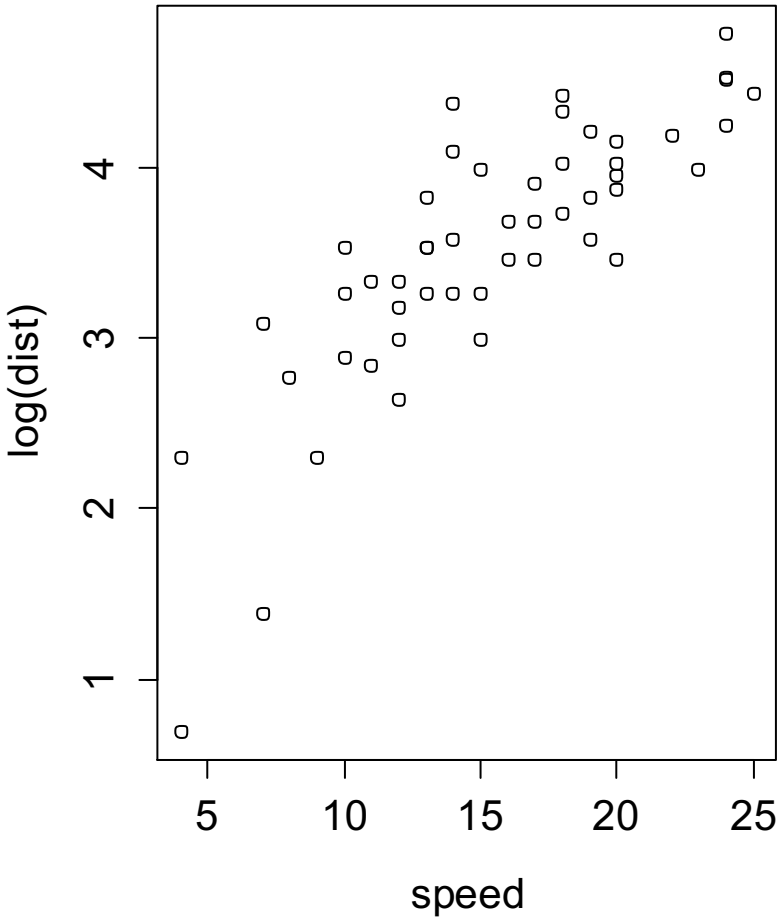


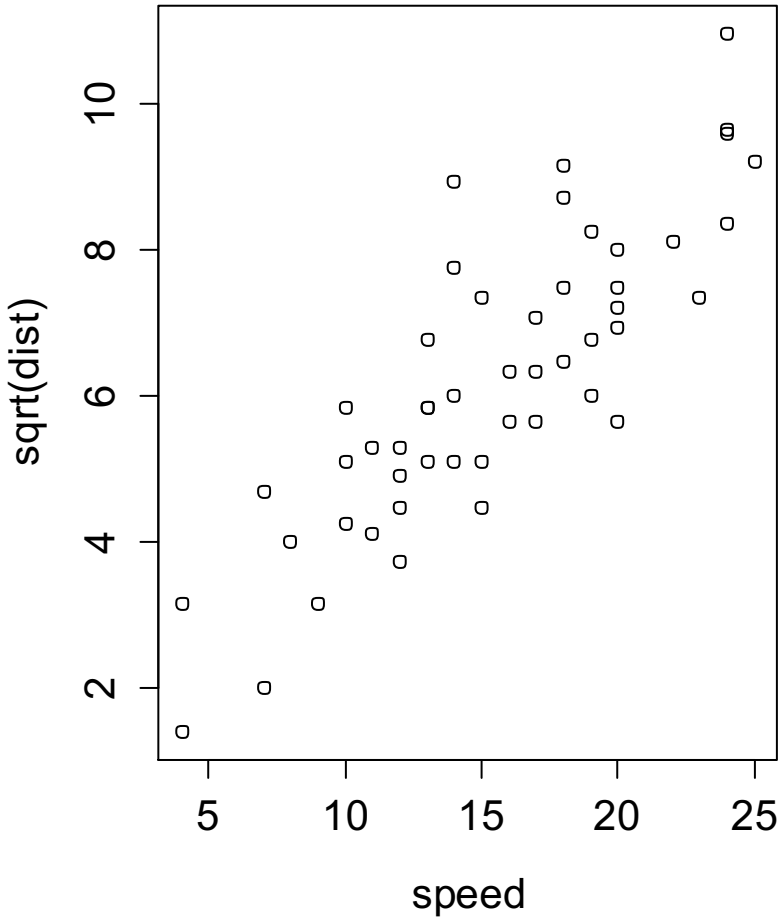
분산이 증가하는 경향 ➔ 종속변수의 log 혹은 sqrt 변환 시도

잔차의 정규성 검정:CARS, No intercept 모형



종속변수 변환






```
> powerTransform(dist~speed+0,data=data)
Estimated transformation parameters
       Y1
0.5039977
```

Sqrt 변환 후 회귀분석: no intercept

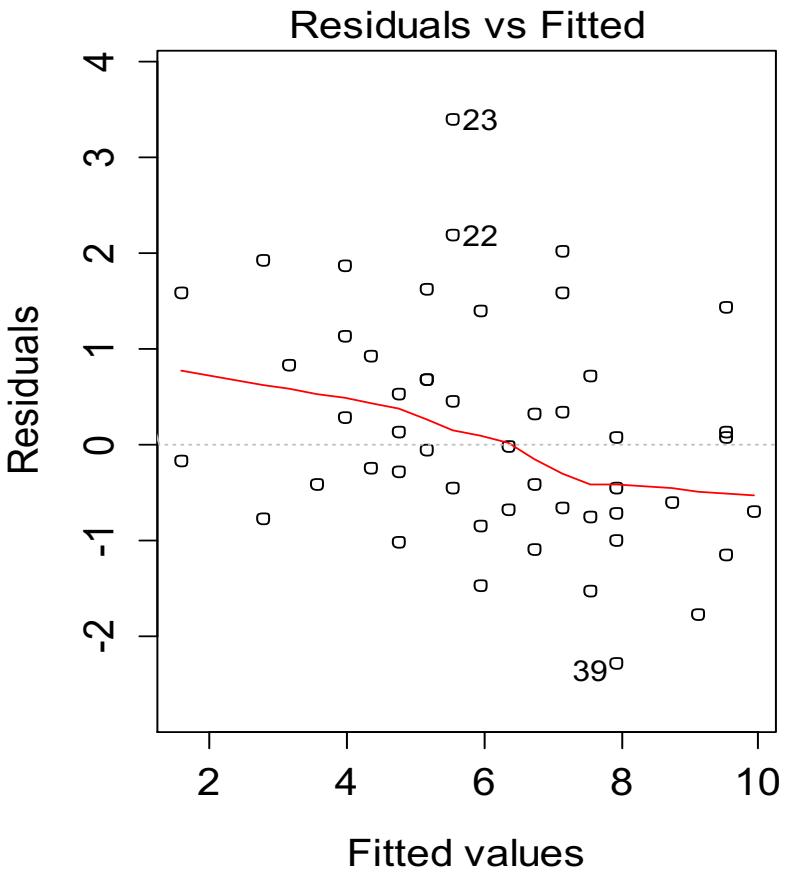
```
Call:
lm(formula = sqrt(dist) ~ speed + 0, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2781 -0.6972  0.0208  0.7965  3.3898

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
speed  0.39675    0.01015   39.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.167 on 49 degrees of freedom
Multiple R-squared: 0.9689,   Adjusted R-squared: 0.9683
F-statistic:  1528 on 1 and 49 DF,  p-value: < 2.2e-16
```

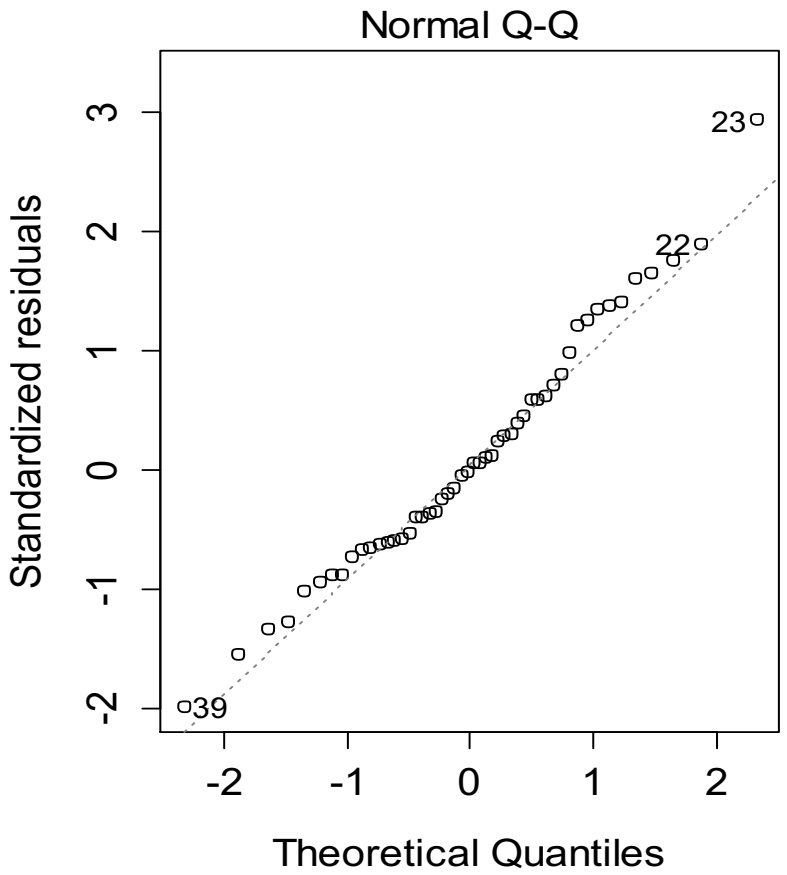
추정된 모형:

기울기 해석: Speed 가 1 증가할 때 _____ 가 _____ 만큼 증가한다.

잔차도, 정규성 검정

```
> out2=lm(sqrt(dist)~speed+0,data=cars)
> plot(out2)
> shapiro.test(resid(out2))

        Shapiro-Wilk normality test

data:  resid(out2)
W = 0.9792, p-value = 0.5185
```





추정과 예측

- 최종모형으로 추정된 회귀식

$$\widehat{\sqrt{dist}} = 0.397 \times speed$$

- 속도가 10 km/hr 또는 30 km/hr 일 때 멈추기 까지 걸린 거리를 예측하면?

```
> new=data.frame(speed=c(10,30))
> predict(out2,new)
        1         2
 3.967494 11.902483
```

신뢰구간, 예측구간

- 속도가 10 또는 30 km/hr 일 때 멈추는데 걸리는 **평균 거리**의 95% 신뢰구간

```
> predict(out2,new,interval="confidence")
        fit       lwr      upr
1  3.967494  3.763518  4.17147
2 11.902483 11.290554 12.51441
```

- **새로운 한 자동차**의 속도가 10 또는 30 km/hr 일 때 멈추는데 걸리는 거리의 95% 예측구간

```
> predict(out2,new,interval="prediction")
        fit      lwr       upr
1  3.967494 1.612650  6.322338
2 11.902483 9.477995 14.326970
```

예측치 (모든 관측치에 대해)

```
> cbind(speed,fitted(out2))
   speed
1      4 1.586998
2      4 1.586998
3      7 2.777246
4      7 2.777246
5      8 3.173995
6      9 3.570745
7     10 3.967494
8     10 3.967494
9     10 3.967494
…
```

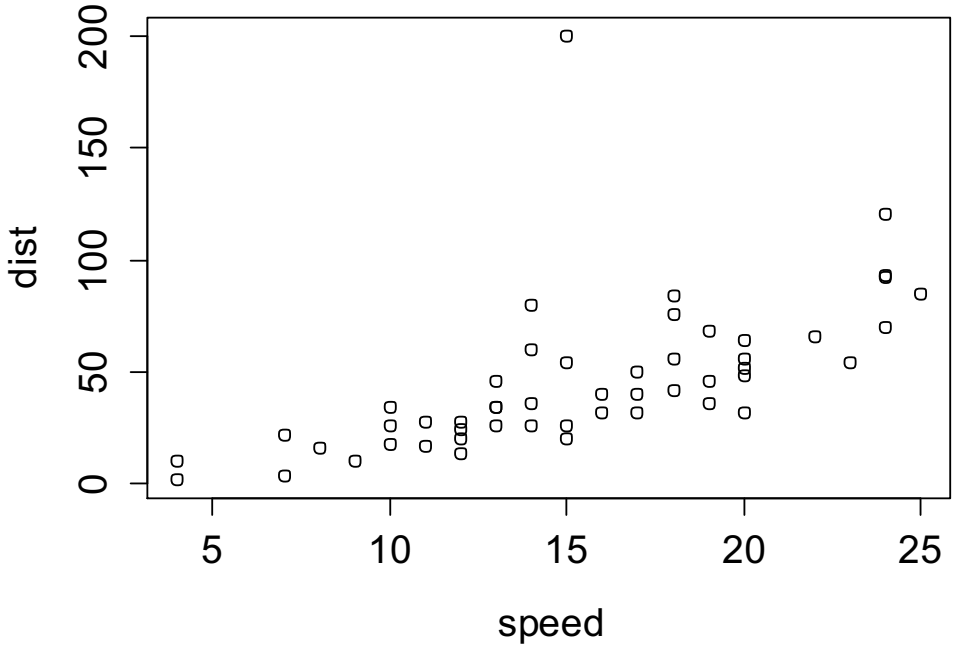
결과 해석의 유의점

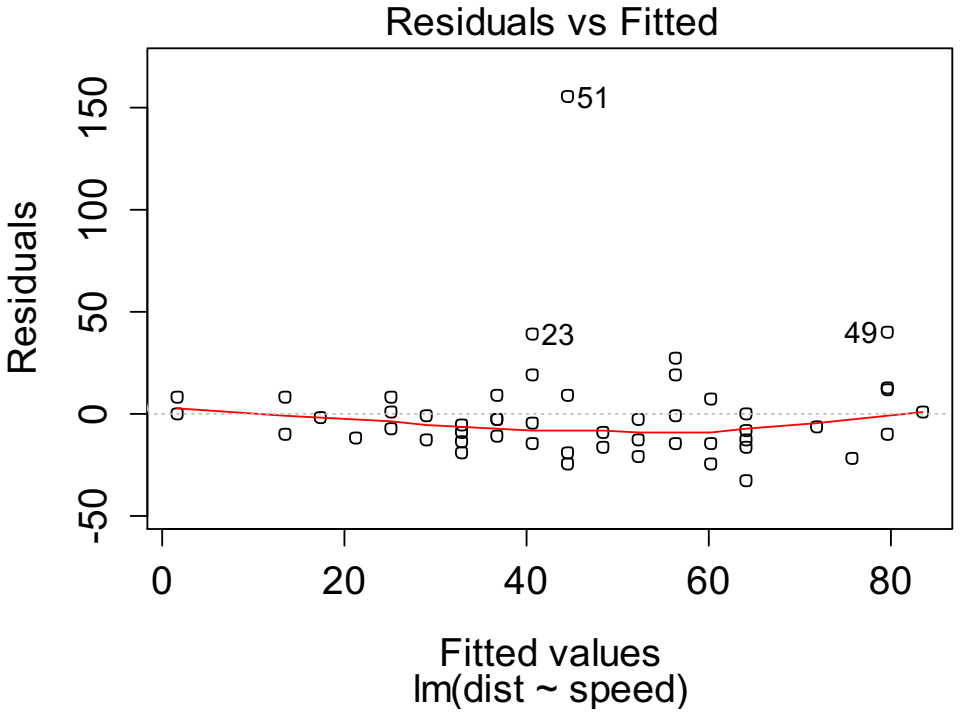
- 회귀식은 가지고 있는 data 범위 밖에서 예측은 주의해야 한다. (Extrapolation 문제)

- 회귀식이 유의하다고 해서 인과관계를 증명하는 것은 아님.

Outlier (이상점) and Influential Points (영향점)

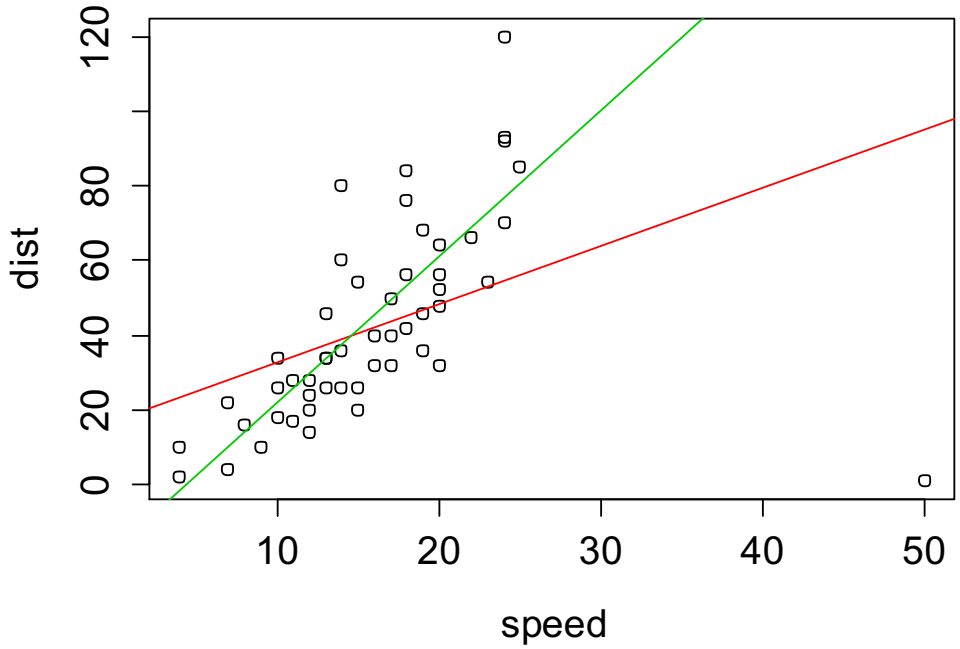
- 이상점
  - 측정상 혹은 실험상의 과오로 조사대상인 모집단에 속하지 않는 다고 의심이 될 정도로 정상범위 밖에 떨어진 점
  - 대개 큰 잔차를 가짐.

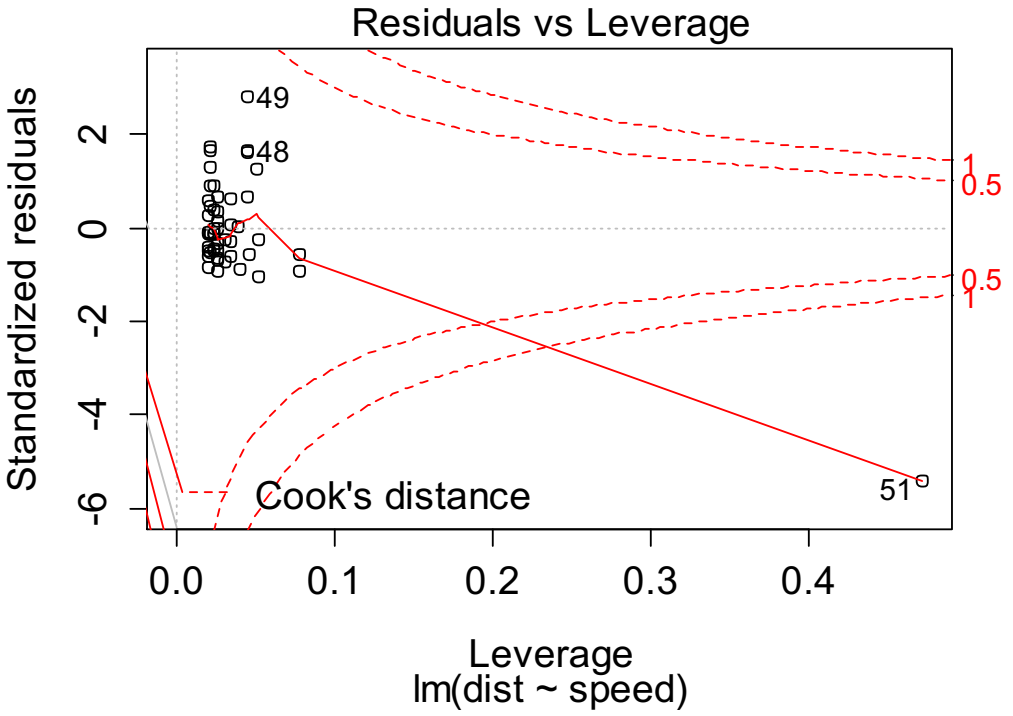




Outlier (이상점) and Influential Points (영향점)

- 영향점
  - 소수의 관측치들이 통계량에 큰 영향





정리: 단순회귀분석의 절차

1. 연구가설 설정
2. 변수탐색
   - 기술통계법 (각 변수의 평균, 표준편차 사례 수 등)
   - 변인 상관관계 분석 (상관계수, 산점도)
   - 필요시 변수변환 (선형관계?)
3. 결정계수, F-test로 모형 유의성 검정
4. 잔차분석 (잔차도, 잔차의 Q-Q plot, Leverage plot)
5. 회귀계수 추정치 분석 및 해석
6. 예측