단순회귀분석

회귀진단

회귀진단

- 만약 회귀모형이 제대로 설정되고 추정되었다면

오차항= 찌꺼기

오차항이 추세를 보인다면 무언가 중요한 정보가 모형에 포함되지 않았다는 의미

오차항에 대한 가정

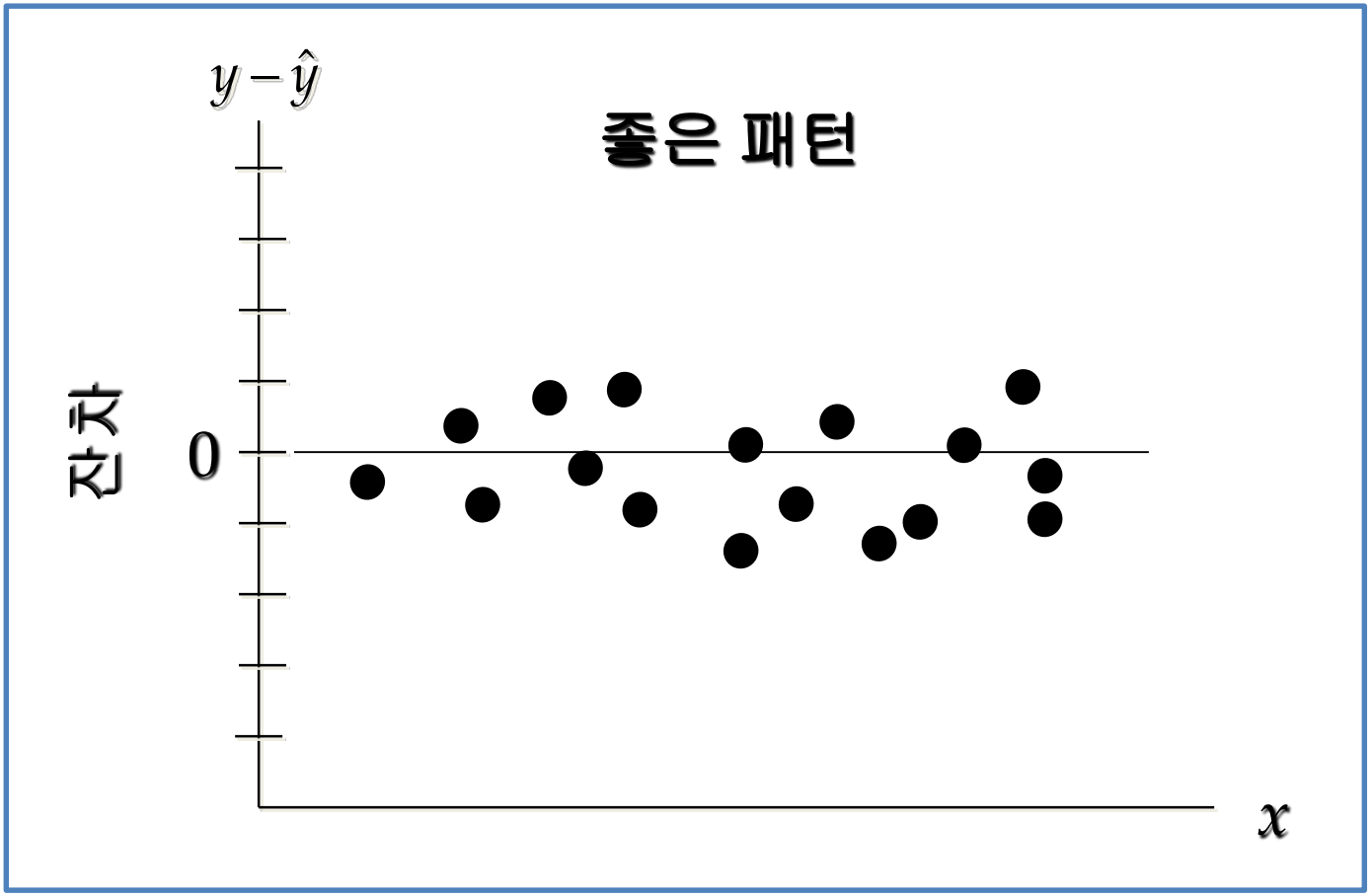
1. 오차항 ε 은 평균이 '0'인 확률변수이다.

2. ε의 분산은 모든 x값에 대해 동일하다.

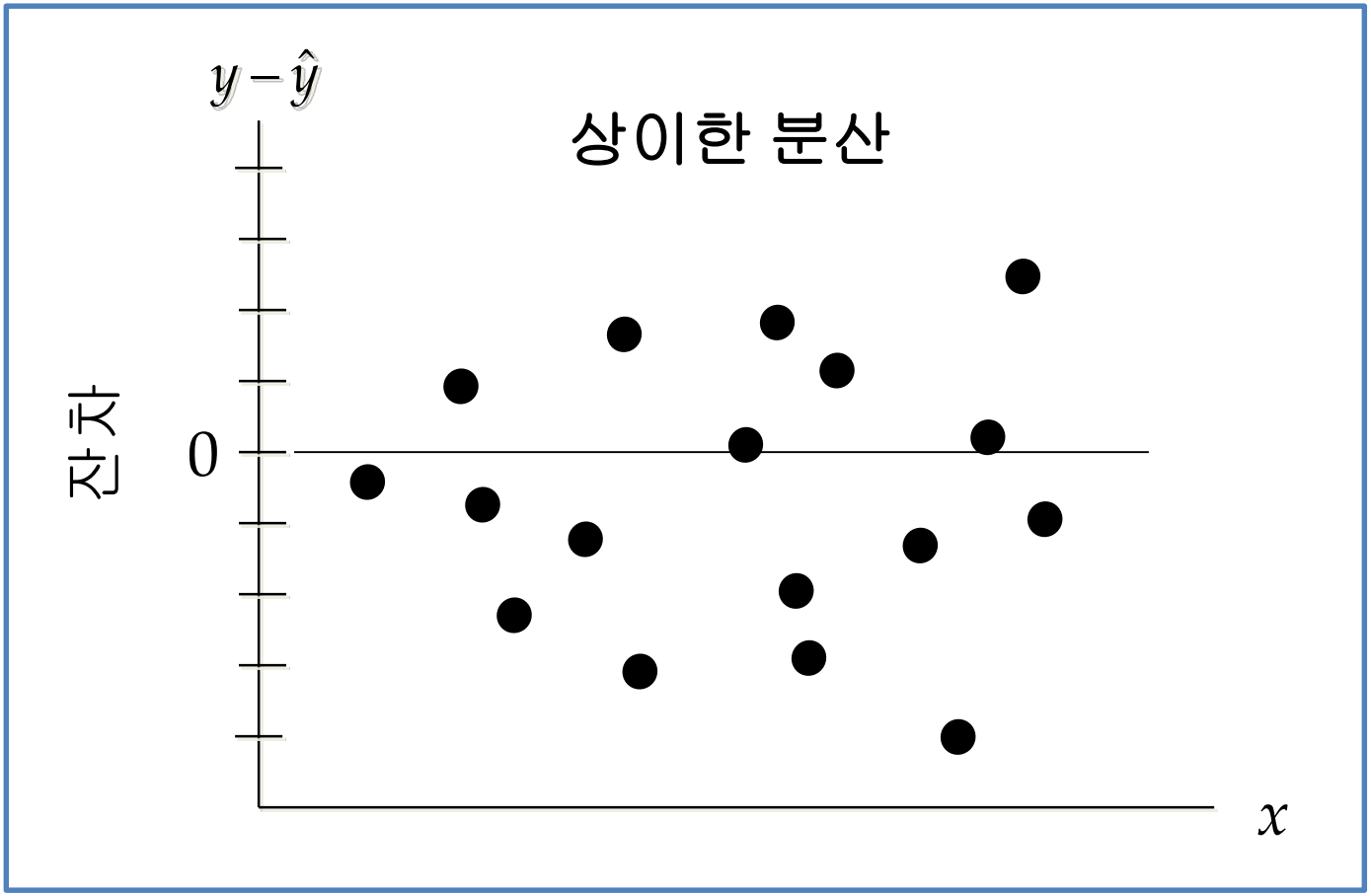
3. ε 값들은 서로 독립적이다.

4. 오차항 ε 은 정규분포를 이루는 확률변수이다.

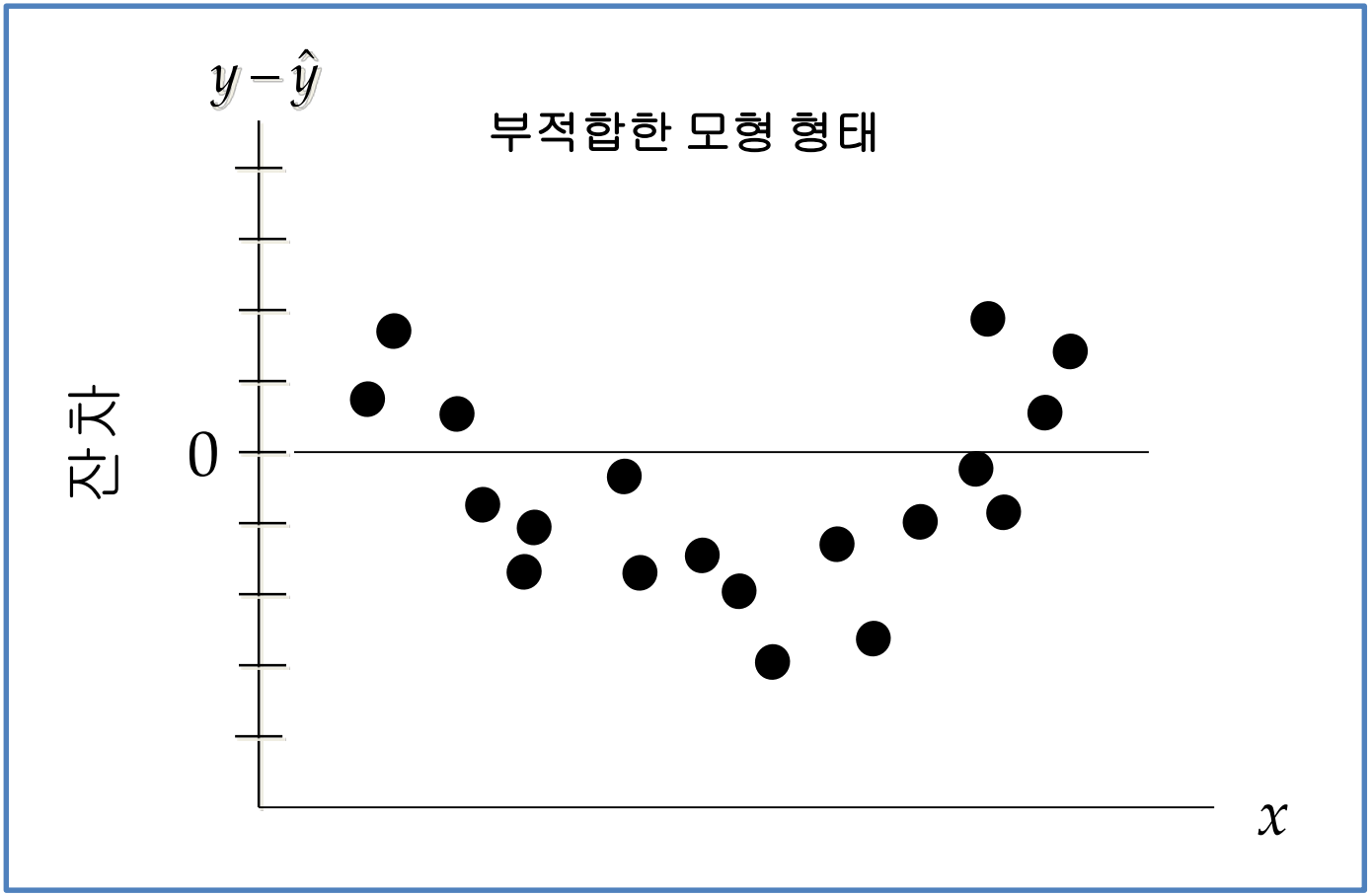
x에 대한 잔차그림



x에 대한 잔차그림

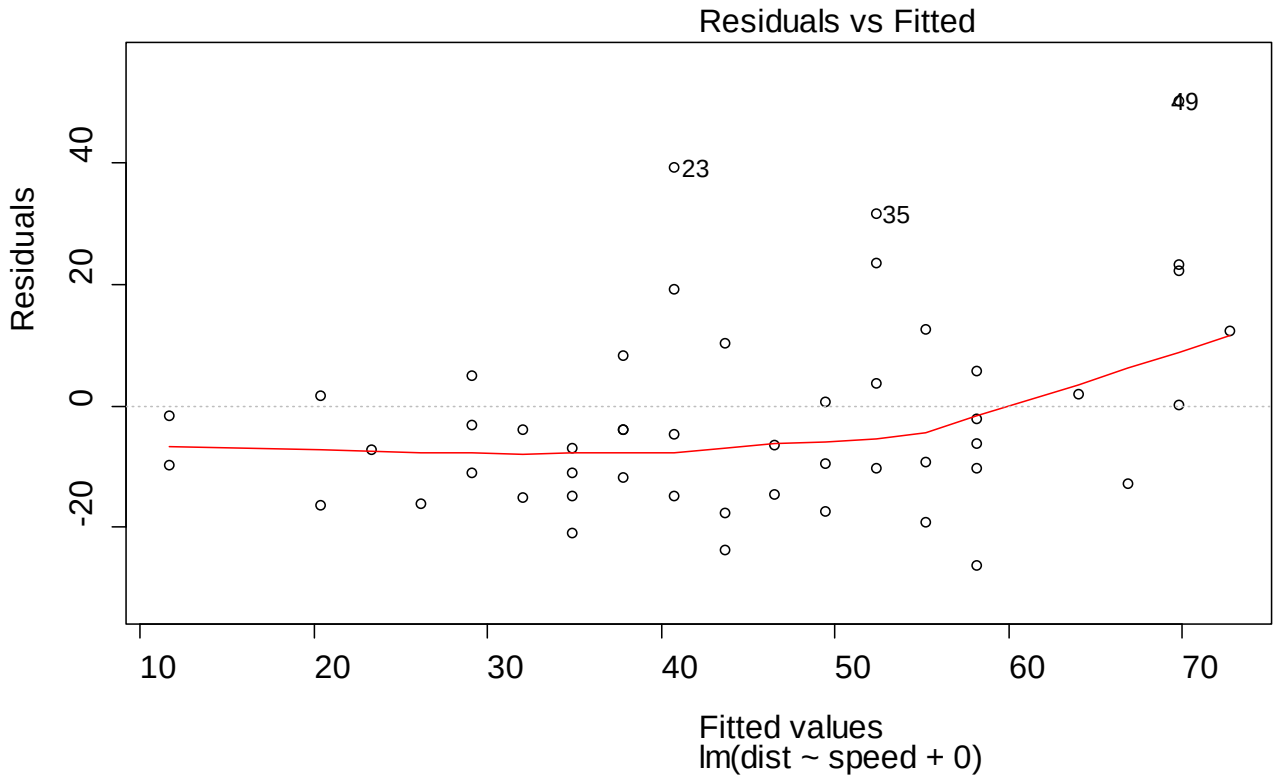


x에 대한 잔차그림



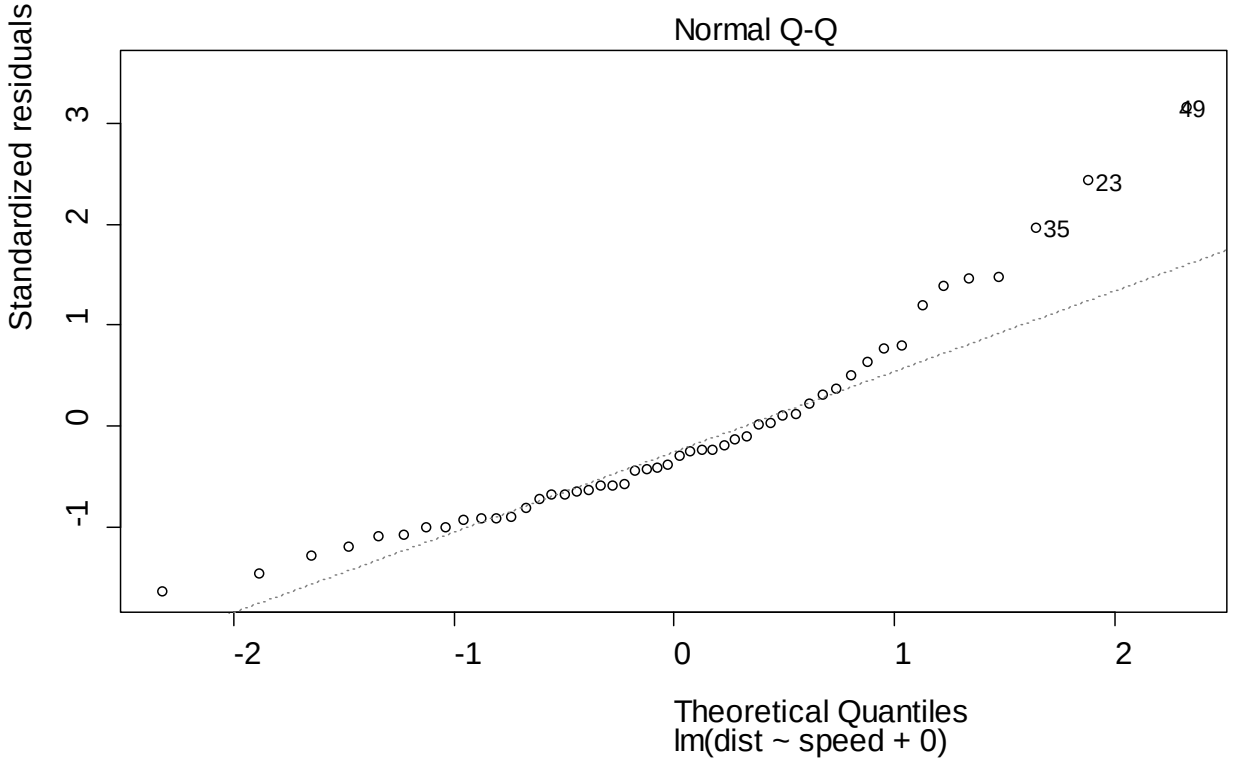
잔차도: CARS, No intercept 모형

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

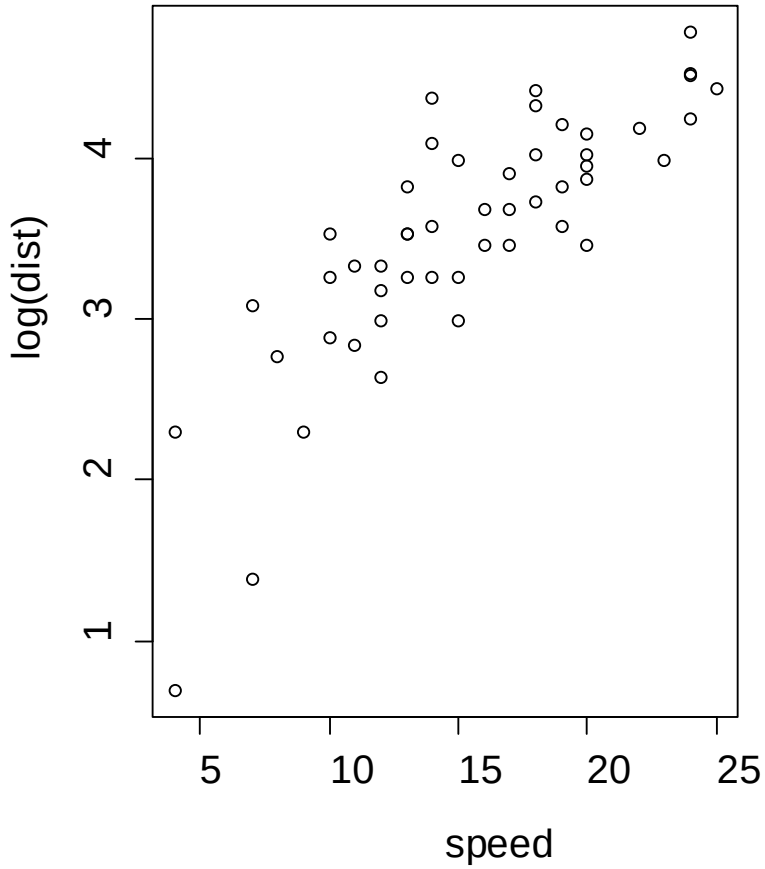


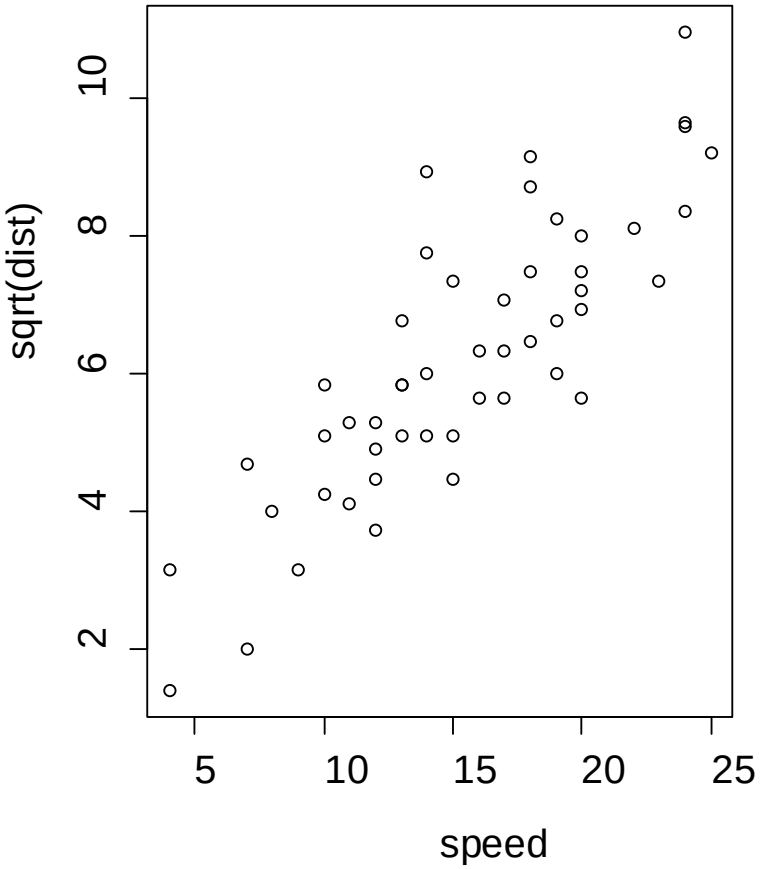
분산이 증가하는 경향 ➔ 종속변수의 log 혹은 sqrt 변환 시도

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:

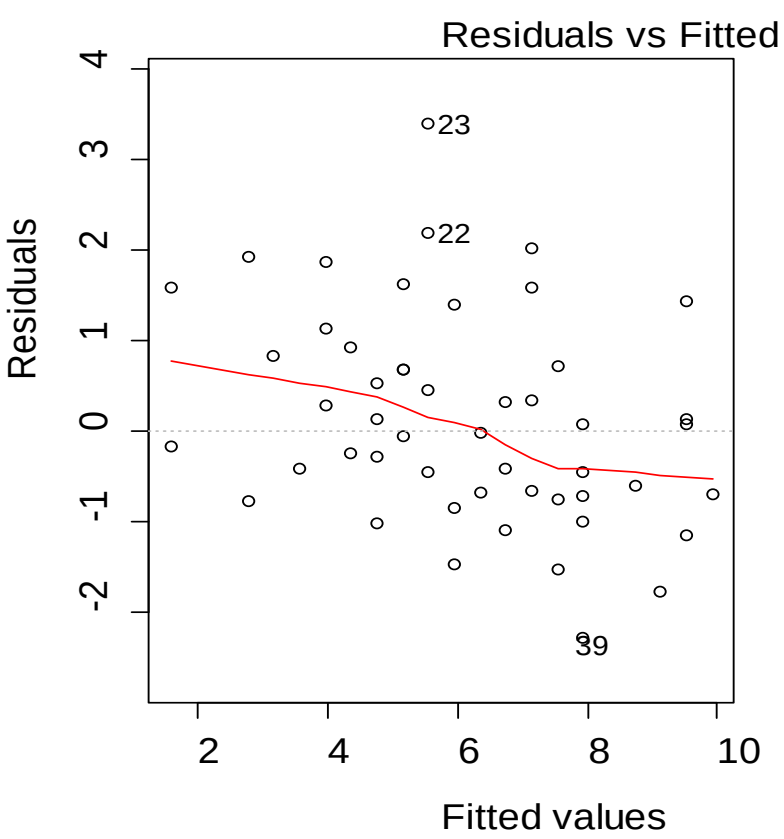
기울기 해석: Speed 가 1 증가할 때 _____ 가 _____ 만큼 증가한다.

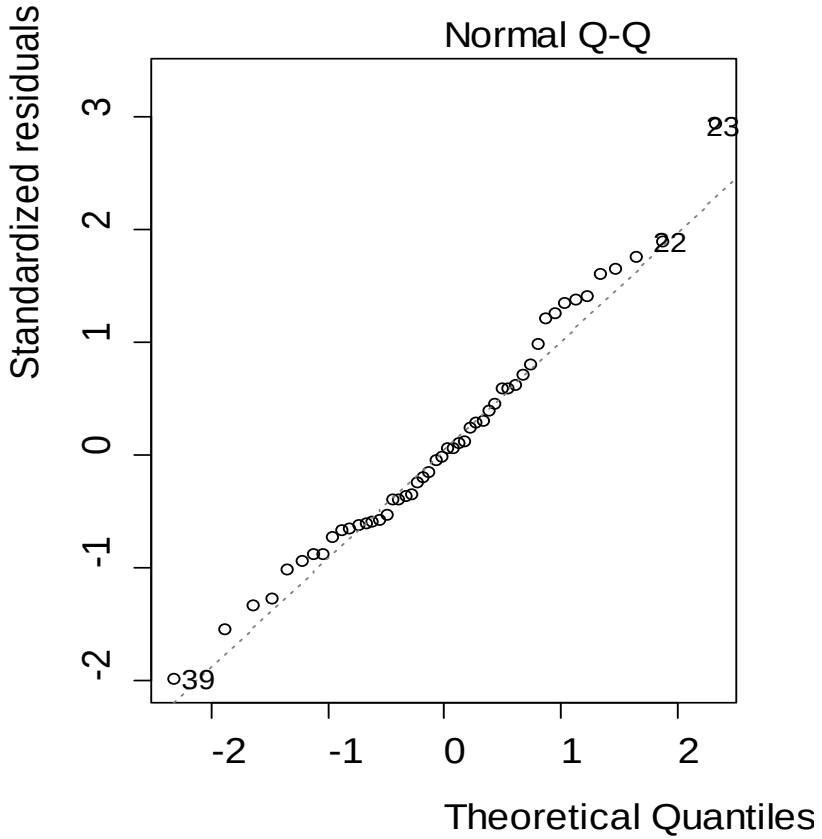
잔차도, 정규성 검정

```
> 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
…
```

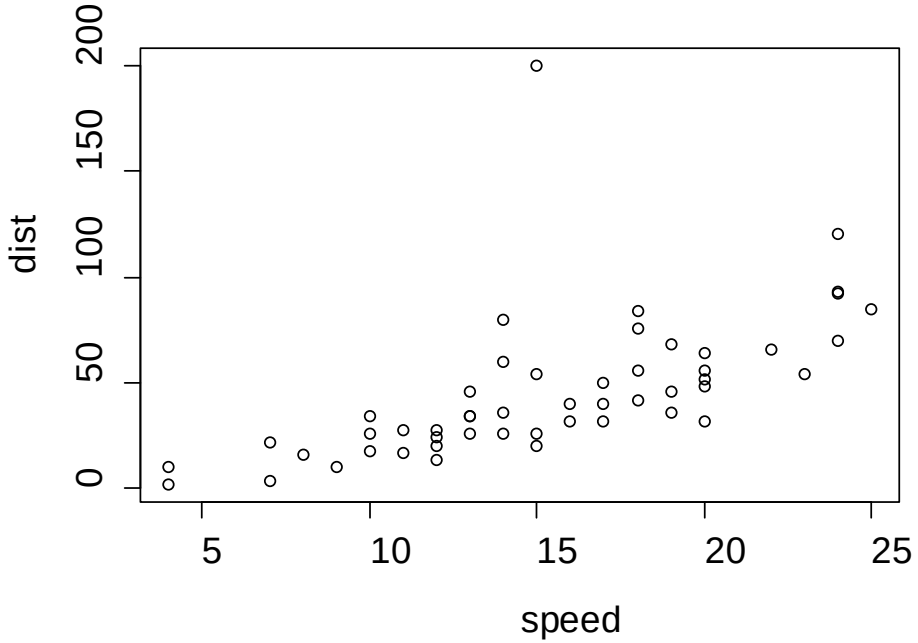
결과 해석의 유의점

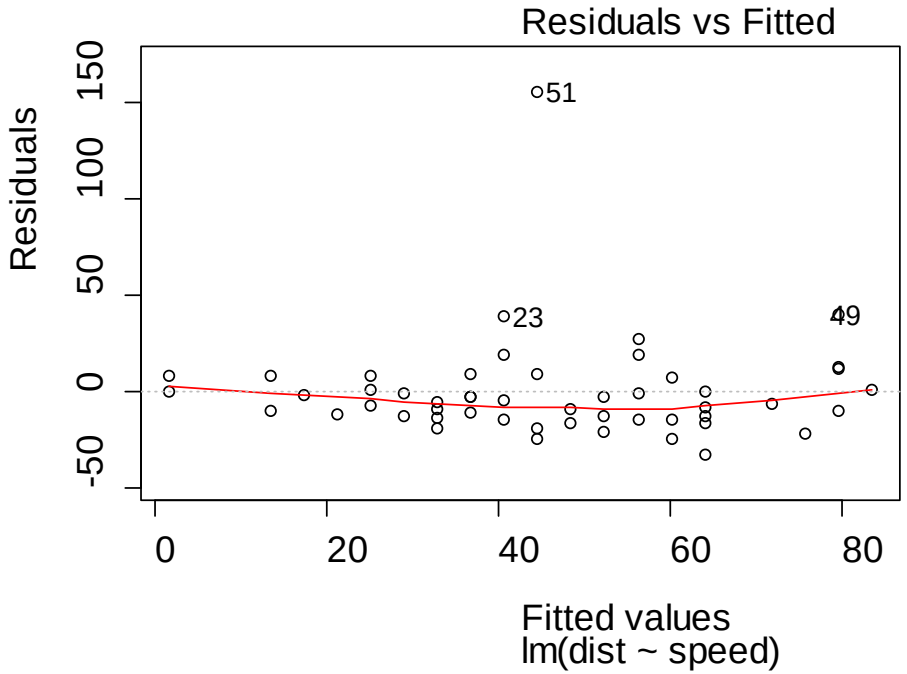
- 회귀식은 가지고 있는 data 범위 밖에서 예측은 주의해야 한다. (Extrapolation 문제)

- 회귀식이 유의하다고 해서 인과관계를 증명하는 것은 아님.

Outlier (이상점) and Influential Points (영향점)

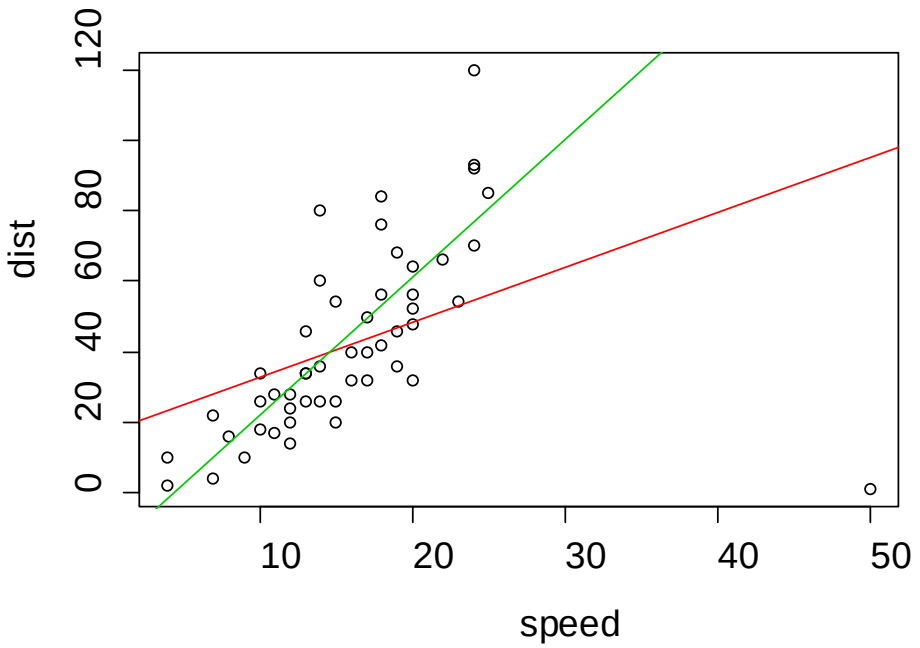
- 이상점
  - 측정상 혹은 실험상의 과오로 조사대상인 모집단에 속하지 않는 다고 의심이 될 정도로 정상범위 밖에 떨어진 점
  - 대개 큰 잔차를 가짐.

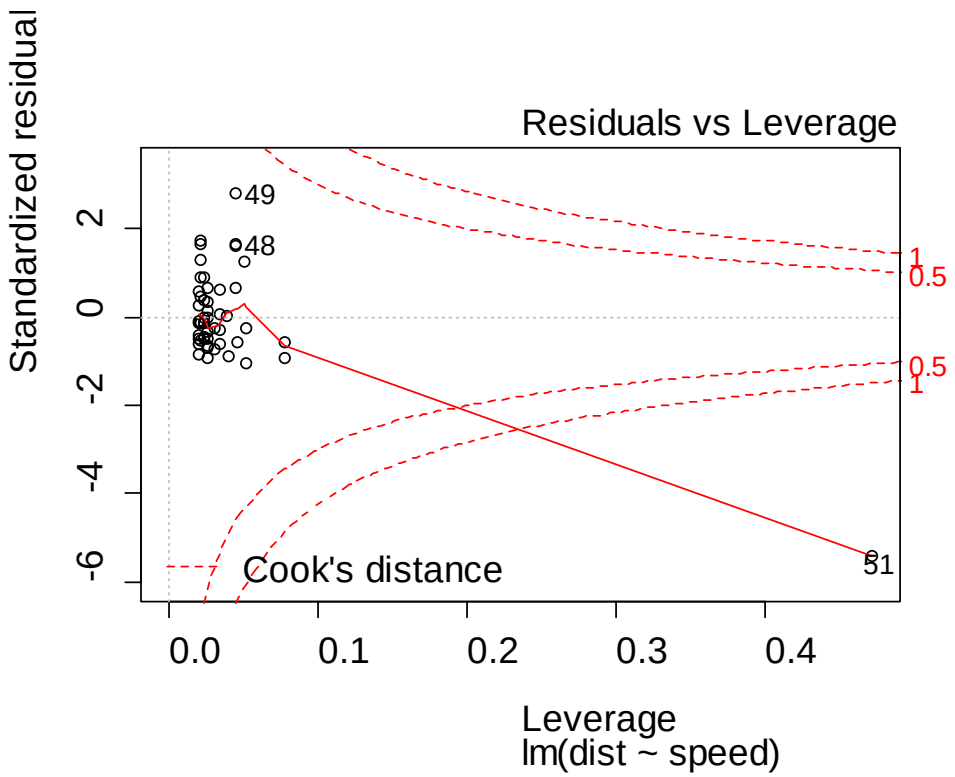




Outlier (이상점) and Influential Points (영향점)

- 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

다중회귀분석

다중회귀모형(multiple regression model)

종속변수 y가 독립변수 x_1, x_2, ...x_p 및 오차항과 어떤 관계가 있는지를 보여주는 식을 다중 회귀모형 이라고 한다.

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p + \varepsilon$$

여기서,

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$는 모수이고,
ε 은 오차항이라 불리는 확률변수이다.

추정 다중회귀식

단순확률표본을 활용하여 모수 $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$의 점추정치인 표본통계량 $b_0, b_1, b_2, \ldots, b_p$를 계산한다.

추정 다중회귀식은 :

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

다중회귀모형

- 예 : Programmer 급여 조사

  한 소프트웨어 회사가 프로그래머 20명에 대한 급여 자료를 수집하였다. 그리고 급여가 경력연수나 직무적성 검사성적과 연관성을 갖는지를 결정하기 위하여 회귀분석이 사용될 수 있다는 제안이 있었다.

경력연수와 직무적성검사 성적과 그에 상응하는 연봉(단위는 $ 천)이 다음 슬라이더에 나타나 있다.

다중회귀모형

경력	점수	연봉	경력	점수	연봉
4	78	24	9	88	38
7	100	43	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29
10	84	38	8	87	34
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30	3	70	28.2
6	91	33	3	89	30

다중회귀모형

연봉 (y)은 경력연수(x_1) 및 직무적성검사 성적(x_2)과 아래와 같은 회귀모형으로 관련되어 있다고 가정한다:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

여기서 ,

y = 연봉 ($1000)

x_1 = 경력연수

x_2 = 직무적성검사 성적

추정된 회귀식

```
> model=lm(salary~experience+score,data)
> summary(model)

Call:
lm(formula = salary ~ experience + score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3586 -1.4581 -0.0341  1.1862  4.9102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17394    6.15607   0.516  0.61279
experience   1.40390    0.19857   7.070 1.88e-06 ***
score        0.25089    0.07735   3.243  0.00478 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.419 on 17 degrees of freedom
Multiple R-squared: 0.8342,     Adjusted R-squared: 0.8147
F-statistic: 42.76 on 2 and 17 DF,  p-value: 2.328e-07
```

SALARY = 3.174 + 1.404(EXPER) + 0.251(SCORE)

계수의 해석방법

다중회귀분석에서 각 회귀계수는 다음과 같이 해석해야 한다.

b_i는 모든 다른 독립변수가 일정할 때 x_i의 1단위 변화에 대한 y값 변화의 추정치이다.

계수의 해석 방법

$b_1 = 1.404$

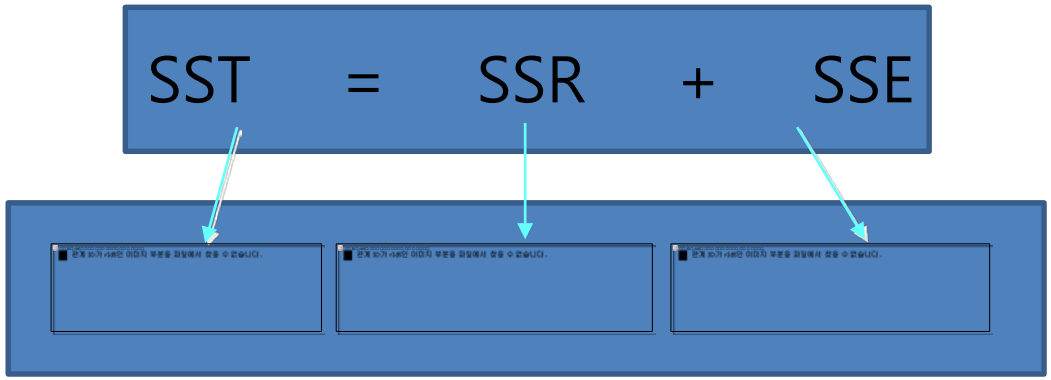
경력 연수가 1년 증가할 때 연봉이 $1,404 증가할 것으로 기대된다 (직무적성검사 성적이 일정 하다고 할 때).

$b_2 = 0.251$

직무적성검사 성적이 1점 올라갈 때 연봉은 $251 올라갈 것으로 기대된다 (경력연수가 일정하다고 할 때).

결정계수 (coefficient of determination; R^2)

- SST, SSR, SSE의 관계



여기서:

SST = 총제곱합
SSR = 회귀제곱합
SSE = 오차제곱합

다중결정계수
(multiple coefficient of determination; R^2)

```
Call:
lm(formula = salary ~ experience + score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3586 -1.4581 -0.0341  1.1862  4.9102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17394    6.15607   0.516  0.61279
experience   1.40390    0.19857   7.070 1.88e-06 ***
score        0.25089    0.07735   3.243  0.00478 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.419 on 17 degrees of freedom
Multiple R-squared: 0.8342,     Adjusted R-squared: 0.8147
F-statistic: 42.76 on 2 and 17 DF,  p-value: 2.328e-07
```

경력연수와 직무적성검사 성적이 연봉의 변동량의 83%를 설명한다

수정 다중결정계수
(adjusted coefficient of determination)

- 설명변수의 수가 증가하면 결정계수는 언제나 증가
- 과연 높은 R^2가 무조건 좋은가? (모수절약의 법칙)
- 설명변수의 개수에 대한 패널티 적용한 결정계수

```
Residual standard error: 2.419 on 17 degrees of freedom
Multiple R-squared: 0.8342,     Adjusted R-squared: 0.8147
F-statistic: 42.76 on 2 and 17 DF,  p-value: 2.328e-07
```

유의성 검정(testing for significance)

단순회귀 분석에서는 *F* 검정과 *t* 검정이 같은 결론을 제공한다.

다중회귀분석에서 *F* 검정의 목적은 *t* 검정의 목적과 다르다 .

- **F검정**
  - *F* 검정은 종속변수와 ***모든 독립변수 집합*** 간에 유의한 관계가 존재하는지를 검정하기 위해 활용된다.

- **T검정**
  - 각 ***개별 독립변수***가 유의한지 여부를 검정하기 위해 활용된다.

유의성 검정: F 검정

가설

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_a: 하나 이상의 모수가 0이 아니다.

```
Call:
lm(formula = salary ~ experience + score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3586 -1.4581 -0.0341  1.1862  4.9102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17394    6.15607   0.516  0.61279
experience   1.40390    0.19857   7.070 1.88e-06 ***
score        0.25089    0.07735   3.243  0.00478 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.419 on 17 degrees of freedom
Multiple R-squared: 0.8342,     Adjusted R-squared: 0.8147
F-statistic: 42.76 on 2 and 17 DF,  p-value: 2.328e-07
```

유의성 검정: t 검정

가설

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

```
Call:
lm(formula = salary ~ experience + score, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3586 -1.4581 -0.0341  1.1862  4.9102

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17394    6.15607   0.516  0.61279
experience   1.40390    0.19857   7.070 1.88e-06 ***
score        0.25089    0.07735   3.243  0.00478 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.419 on 17 degrees of freedom
Multiple R-squared: 0.8342,     Adjusted R-squared: 0.8147
F-statistic: 42.76 on 2 and 17 DF,  p-value: 2.328e-07
```

오차항에 대한 가정

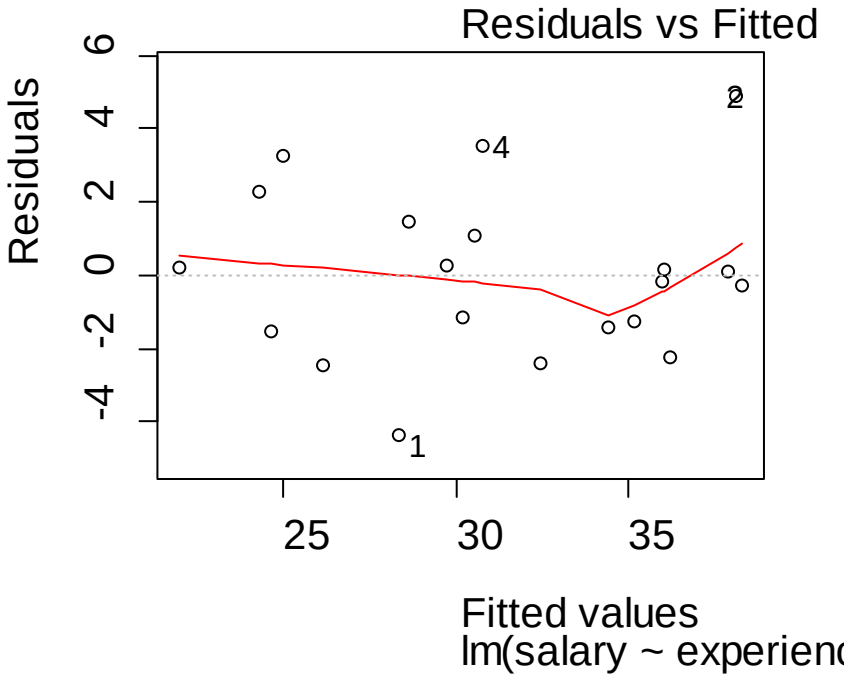
1. 오차항 ε 은 평균이 '0'인 확률변수이다.

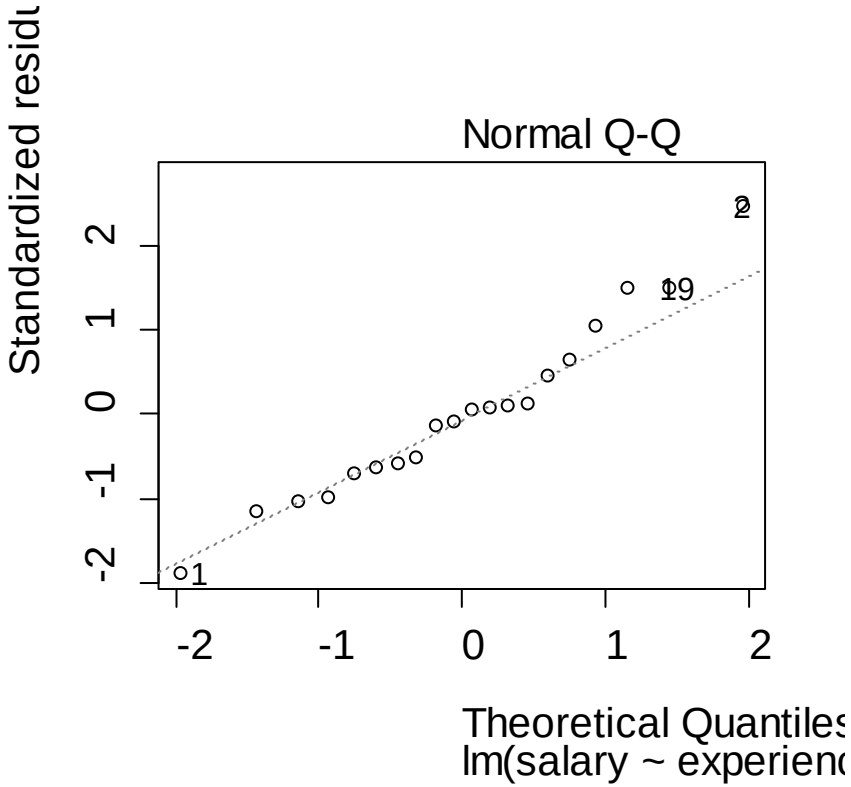
2. ε의 분산은 모든 x값에 대해 동일하다.

3. ε 값들은 서로 독립적이다.

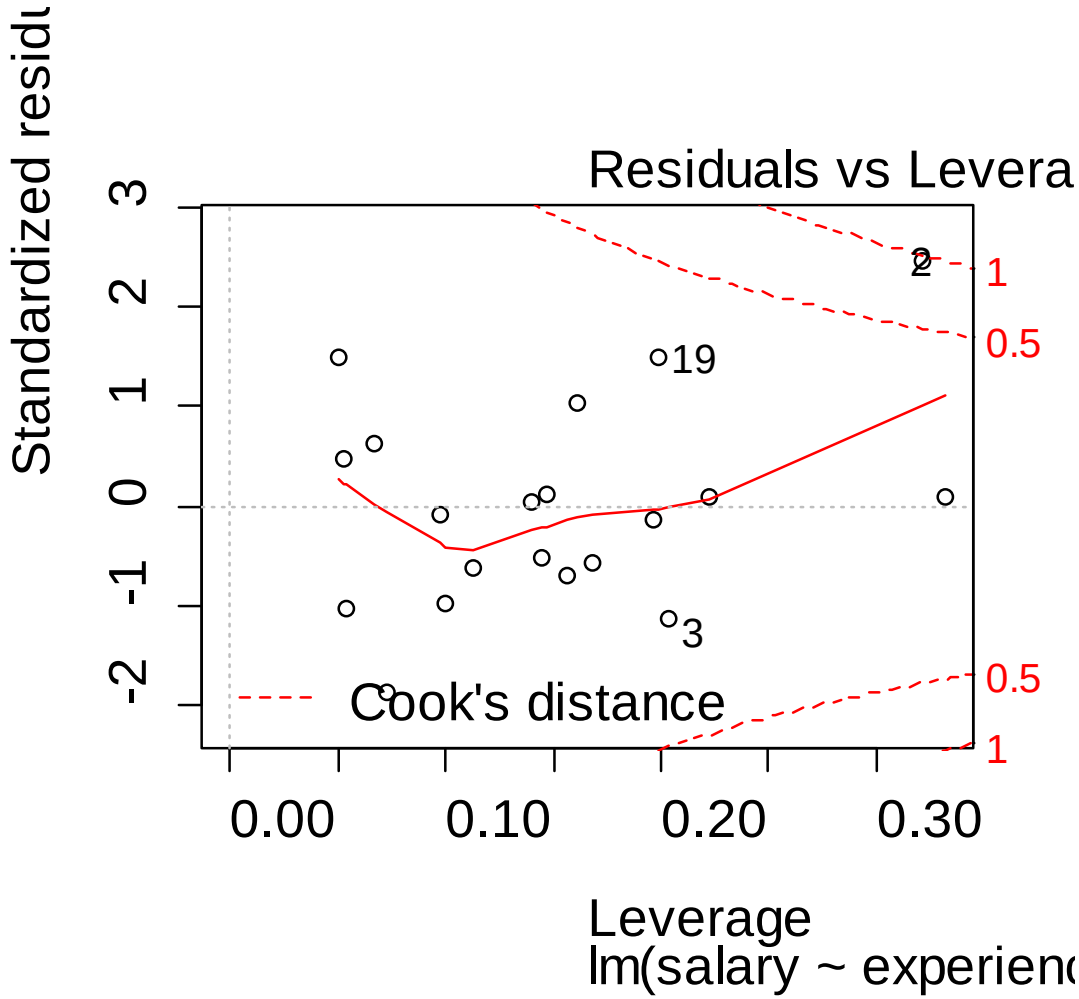
4. 오차항 ε 은 정규분포를 이루는 확률변수이다.

잔차분석


Residuals vs Fitted
Residuals
Fitted values
lm(salary ~ experienc
25
30
35
-4
-2
0
2
4
6
1
4
2


Normal Q-Q
Standardized residu
Theoretical Quantiles
lm(salary ~ experienc
-2
-1
0
1
2
1
19
2

잔차분석


Standardized residu
Residuals vs Levera
2
19
3
1
0.5
0.5
1
Cook's distance
-2
-1
0
1
2
3
0.00
0.10
0.20
0.30
Leverage
lm(salary ~ experienc

추정과 예측

- 경력 5년, 적성검사 성적 80점인 사람과 경력 10년, 적성검사 성적 70점인 사람의 연봉 예측치는?

```
> 
predict(model,data.frame("experience"=c(5,10),"score"=c(80
,70)))
       1        2 
30.26428 34.77494 
```

유의성 검정: 다중공선성(multicollinearity)

다중공선성은 독립변수들 사이의 상관관계를 지칭한다.

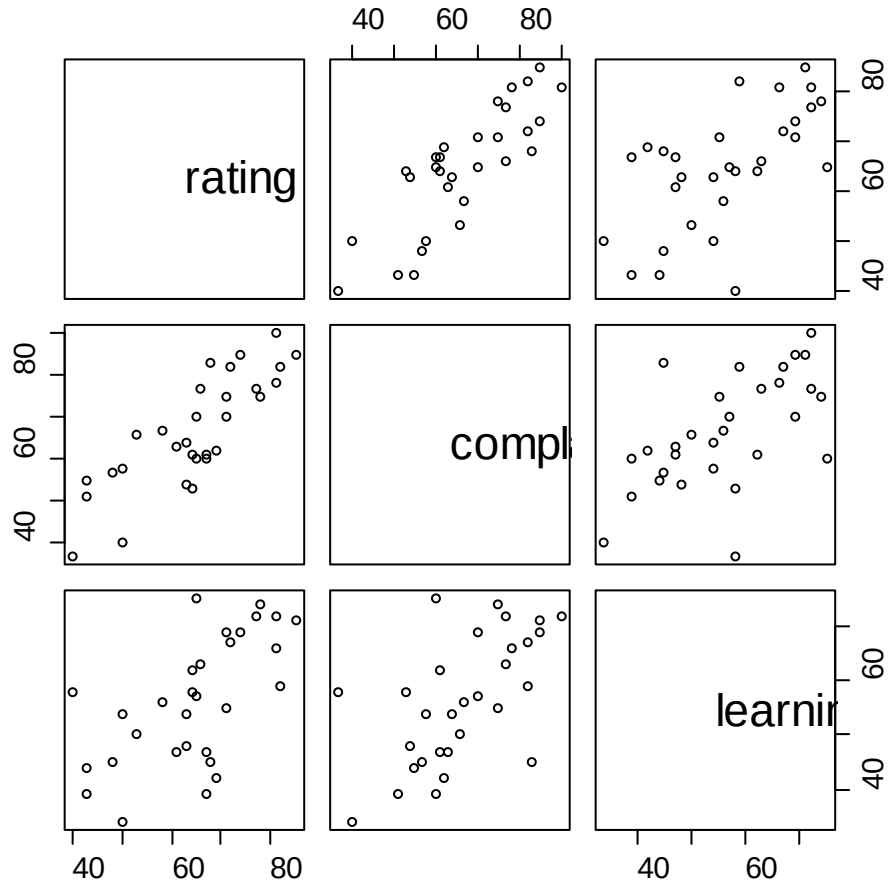
독립변수들이 높은 상관관계를 가질 때,
어떤 특정 독립변수가 종속변수에 미치는 개별적인 영향을 파악하기 어렵다.

유의성 검정: 다중공선성(multicollinearity)

Attitude 자료

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

유의성 검정: 다중공선성(multicollinearity)



- 상관계수

```
             rating complaints  learning
rating     1.0000000  0.8254176 0.6236782
complaints 0.8254176  1.0000000 0.5967358
learning   0.6236782  0.5967358 1.0000000
```

유의성 검정: 다중공선성(multicollinearity)

```
> summary(lm(rating~complaints+learning,data=attitude))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571    0.128
```

Learning이 1 증가할 때 rating이 0.2112만큼 증가한다고 기대한다. **(complaints가 일정하게 유지될 때)**

Learning은 rating을 설명하기에 유의하지 않은 변수인가?

```
> summary(lm(rating~learning,data=attitude))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1741     8.8148   3.196 0.003438 **
learning      0.6468     0.1532   4.222 0.000231 ***
```

모형선택 (Model Selection)

- Confirmatory Analysis
  - 모형선택이 이론에 근거를 둔 경우

- Exploratory Analysis
  - 적용할 이론을 사전에 정해놓지 않고 가능한 여러 모형을 고려한 후 가장 적절한 모형을 고르는 분석
  - 모형선택 방법을 통해 독립변수의 수를 줄인다.

모형선택 (Model Selection)

- Confirmatory Analysis
  - 모형선택이 이론에 근거를 둔 경우

- Exploratory Analysis
  - 적용할 이론을 사전에 정해놓지 않고 가능한 여러 모형을 고려한 후 가장 적절한 모형을 고르는 분석
  - 모형선택 방법을 통해 독립변수의 수를 줄인다.

모형선택 방법

- Forward selection
  - 가장 유의한 변수부터 하나씩 추가
- Backward selection
  - 모든 변수를 넣고 가장 기여도가 낮은 것부터 하나씩 제거
- Stepwise selection
  - Forward selection과 backward selection을 조합

- All subsets
  - 모든 가능한 모형 을 비교하여 최적의 모형 선택
  - 여러 모형 중 최소 AIC, BIC, Mallow's Cp 혹은 최대 adjusted R-sq를 갖는 모형을 선택

Backward Selection

- 모든 변수를 넣고 모형을 추정한다.

```
> out=lm(rating~.,data=attitude)
> anova(out)
Analysis of Variance Table

Response: rating
           Df  Sum Sq Mean Sq F value    Pr(>F)
complaints  1 2927.58 2927.58 58.6026 9.056e-08 ***
privileges  1    7.52    7.52  0.1505    0.7016
learning    1  137.25  137.25  2.7473    0.1110
raises      1    0.94    0.94  0.0189    0.8920
critical    1    0.56    0.56  0.0113    0.9163
advance     1   74.11   74.11  1.4835    0.2356
Residuals  23 1149.00   49.96
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

가장 유의하지 않은 변수를 제거

Backward Selection

- 가장유의하지 않은 변수 하나 제거후 다시 모형 추정

```
> out2=lm(rating~.-critical,data=attitude)
> anova(out2)
Analysis of Variance Table

Response: rating
           Df  Sum Sq Mean Sq F value    Pr(>F)
complaints  1 2927.58 2927.58 60.9698 4.835e-08 ***
privileges  1    7.52    7.52  0.1566    0.6958
learning    1  137.25  137.25  2.8583    0.1039
raises      1    0.94    0.94  0.0196    0.8898
advance     1   71.27   71.27  1.4842    0.2350
Residuals  24 1152.41   48.02
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Backward Selection

```
> backward=step(out,direction="backward",trace=FALSE)
> backward$anova
          Step Df  Deviance Resid. Df Resid. Dev      AIC
1              NA        NA        23   1149.000 123.3635
2   - critical  1  3.405864        24   1152.406 121.4523
3     - raises  1 10.605443        25   1163.012 119.7271
4 - privileges  1 16.097508        26   1179.109 118.1395
5    - advance  1 75.539831        27   1254.649 118.0024
> summary(backward)

Call:
lm(formula = rating ~ complaints + learning, data = attitude)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5568  -5.7331   0.6701   6.5341  10.3610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
complaints    0.6435     0.1185   5.432 9.57e-06 ***
learning      0.2112     0.1344   1.571    0.128
---
```

Stepwise Selection

```
> both=step(out,direction="both",trace=FALSE)
> both$anova
          Step Df  Deviance Resid. Df Resid. Dev      AIC
1               NA        NA        23   1149.000 123.3635
2   - critical  1  3.405864        24   1152.406 121.4523
3     - raises  1 10.605443        25   1163.012 119.7271
4 - privileges  1 16.097508        26   1179.109 118.1395
5    - advance  1 75.539831        27   1254.649 118.0024
```

All Subsets Regression

Full model

```
library(leaps)
leaps=regsubsets(rating~.,data=attitude,nbest=5)
```

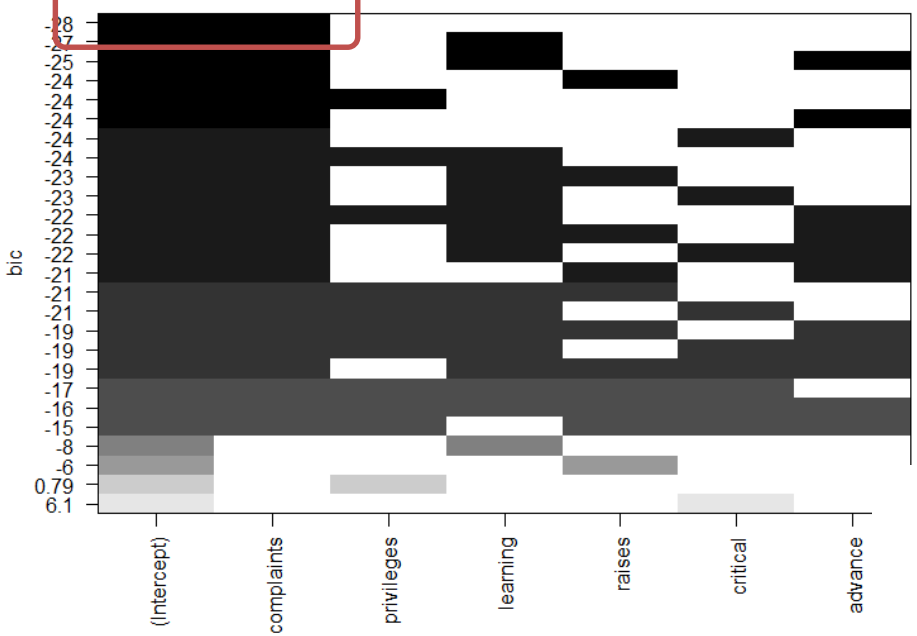
subset의 각 size 당 몇 개의 최적 모형을 저장할 것인가 설정

```
> summary(leaps)
Subset selection object
Call: regsubsets.formula(rating ~ ., data = attitude, nbest = 5)
6 Variables  (and intercept)
           Forced in Forced out
complaints     FALSE      FALSE
privileges     FALSE      FALSE
learning       FALSE      FALSE
raises         FALSE      FALSE
critical       FALSE      FALSE
advance        FALSE      FALSE
5 subsets of each size up to 6
Selection Algorithm: exhaustive
         complaints privileges learning raises critical advance
1  ( 1 ) "*"        " "        " "      " "    " "      " "
1  ( 2 ) " "        " "        "*"      " "    " "      " "
1  ( 3 ) " "        " "        " "      "*"    " "      " "
1  ( 4 ) " "        "*"        " "      " "    " "      " "
1  ( 5 ) " "        " "        " "      " "    "*"      " "
2  ( 1 ) "*"        " "        "*"      " "    " "      " "
2  ( 2 ) "*"        " "        " "      "*"    " "      " "
2  ( 3 ) "*"        "*"        " "      " "    " "      " "
2  ( 4 ) "*"        " "        " "      " "    " "      "*"
2  ( 5 ) "*"        " "        " "      " "    "*"      " "
3  ( 1 ) "*"        " "        "*"      " "    " "      "*"
3  ( 2 ) "*"        "*"        "*"      " "    " "      " "
3  ( 3 ) "*"        " "        "*"      "*"    " "      " "
3  ( 4 ) "*"        " "        "*"      " "    "*"      " "
3  ( 5 ) "*"        " "        " "      "*"    " "      "*"
4  ( 1 ) "*"        "*"        "*"      " "    " "      "*"
4  ( 2 ) "*"        " "        "*"      "*"    " "      "*"
4  ( 3 ) "*"        " "        "*"      " "    "*"      "*"
4  ( 4 ) "*"        "*"        "*"      "*"    " "      " "
4  ( 5 ) "*"        "*"        "*"      " "    "*"      " "
5  ( 1 ) "*"        "*"        "*"      "*"    " "      "*"
5  ( 2 ) "*"        "*"        "*"      " "    "*"      "*"
5  ( 3 ) "*"        " "        "*"      "*"    "*"      "*"
5  ( 4 ) "*"        "*"        "*"      "*"    "*"      " "
5  ( 5 ) "*"        "*"        " "      "*"    "*"      "*"
6  ( 1 ) "*"        "*"        "*"      "*"    "*"      "*"
```

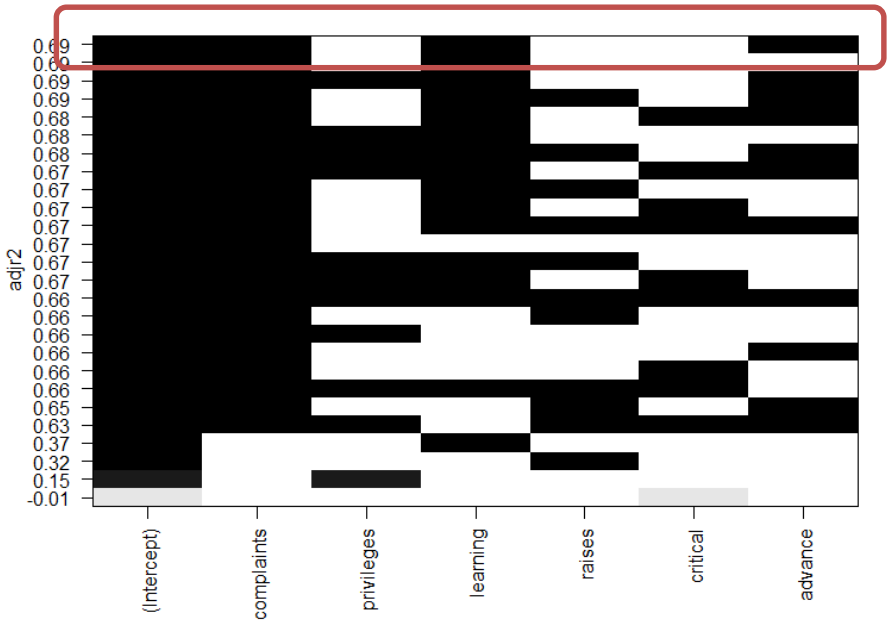
저장된 각 모형에 포함된 설명변수 표시

```
plot(leaps)
plot(leaps,scale="adjr2")
plot(leaps,scale="Cp")
```

Best model



Best model



- adjusted R-square가 최대인 Best model

```
> summary.out=summary(leaps)
> which.max(summary.out$adjr2)
[1] 11
> summary.out$which[11,]
(Intercept)  complaints  privileges    learning      raises    critical     advance
       TRUE        TRUE       FALSE        TRUE       FALSE       FALSE        TRUE
```

```
> out3=lm(rating~complaints+learning+advance,data=attitude)
> summary(out3)

Call:
lm(formula = rating ~ complaints + learning + advance, data = attitude)

Residuals:
    Min      1Q  Median      3Q     Max
-12.217  -5.377   0.967   6.078  11.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5777     7.5439   1.800   0.0835 .
complaints    0.6227     0.1181   5.271 1.65e-05 ***
learning      0.3124     0.1542   2.026   0.0532 .
advance      -0.1870     0.1449  -1.291   0.2082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.734 on 26 degrees of freedom
Multiple R-squared:  0.7256, Adjusted R-squared:  0.6939
F-statistic: 22.92 on 3 and 26 DF,  p-value: 1.807e-07
```