

다중회귀분석

다중회귀모형(multiple regression model)

종속 변수 y 가 독립 변수 x_1, x_2, \dots, x_p 과 어떤 관계가 있는지를 보여주는 식

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$: 회귀계수
- ϵ : 오차항

추정 다중회귀식: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$

예 : Programmer 급여 조사

한 소프트웨어 회사가 프로그래머 20명에 대한 급여 자료를 수집하였다. 그리고 급여가 경력연수나 직무적성 검사성적과 연관성을 갖는지를 결정하기 위하여 회귀분석이 사용하기로 했다.



```
> salary=read.csv("salary.csv")
> head(salary)
  experience score salary
1          4    78   24.0
2          7   100   43.0
3          1    86   23.7
4          5    82   34.3
5          8    86   35.8
6         10    84   38.0
> summary(salary)
      experience          score          salary
Min.   : 0.00   Min.   : 70.00   Min.   :22.20
1st Qu.: 3.00   1st Qu.: 77.25   1st Qu.:27.80
Median : 5.50   Median : 82.50   Median :30.85
Mean   : 5.20   Mean   : 82.75   Mean   :31.23
3rd Qu.: 7.25   3rd Qu.: 87.25   3rd Qu.:34.67
Max.   :10.00   Max.   :100.00   Max.   :43.00
```

다중회귀모형



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

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

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

$$\text{SALARY} = 3.174 + 1.404(\text{EXPER}) + 0.251(\text{SCORE})$$

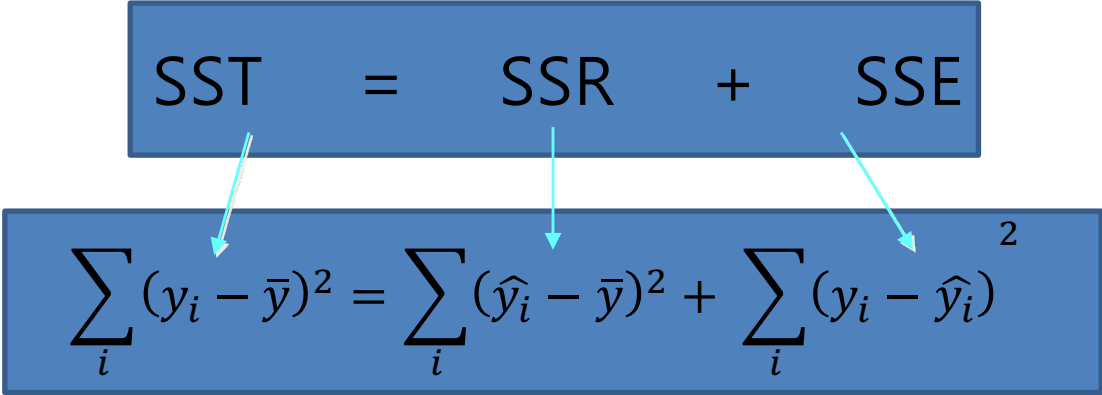
계수의 해석방법

- b_i 는 모든 다른 독립변수가 일정할 때 x_i 의 1단위 변화에 대한 y 값 변화의 추정치
- $b_1 = 1.404$
 - 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 (R^2)

■ SST, SSR, SSE의 관계

$$\text{SST} = \text{SSR} + \text{SSE}$$


$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

여기서:

SST = 총제곱합

SSR = 회귀제곱합

SSE = 오차제곱합

```

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 R²)



$$R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

- 설명변수의 수가 증가하면 결정계수는 언제나 증가
- 과연 높은 R²가 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 .

➤ F 검정

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

➤ T 검정

- 각 개별 독립변수가 유의한지 여부를 검정하기 위해 활용된다.

유의성 검정: F 검정

가설

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

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

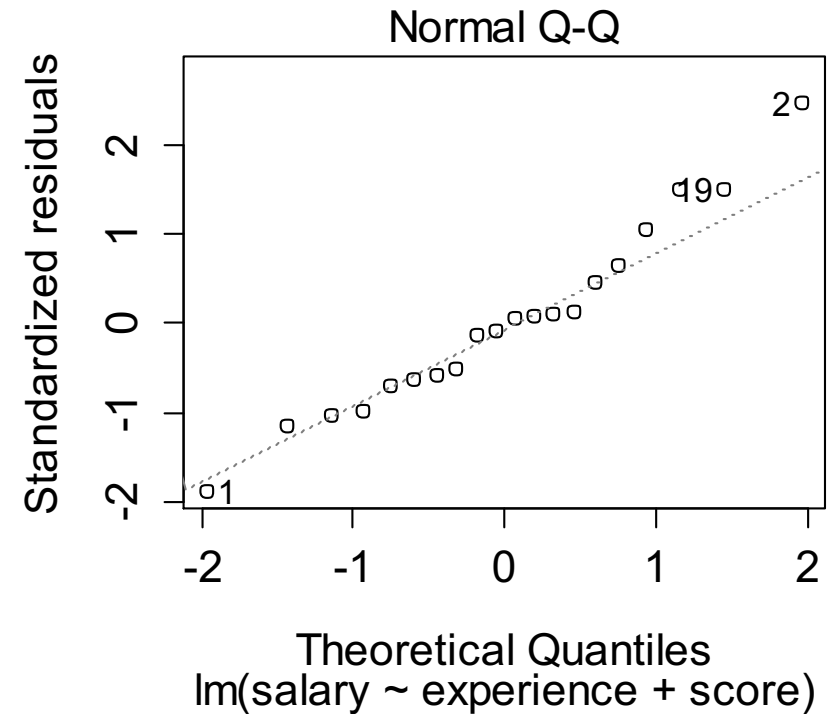
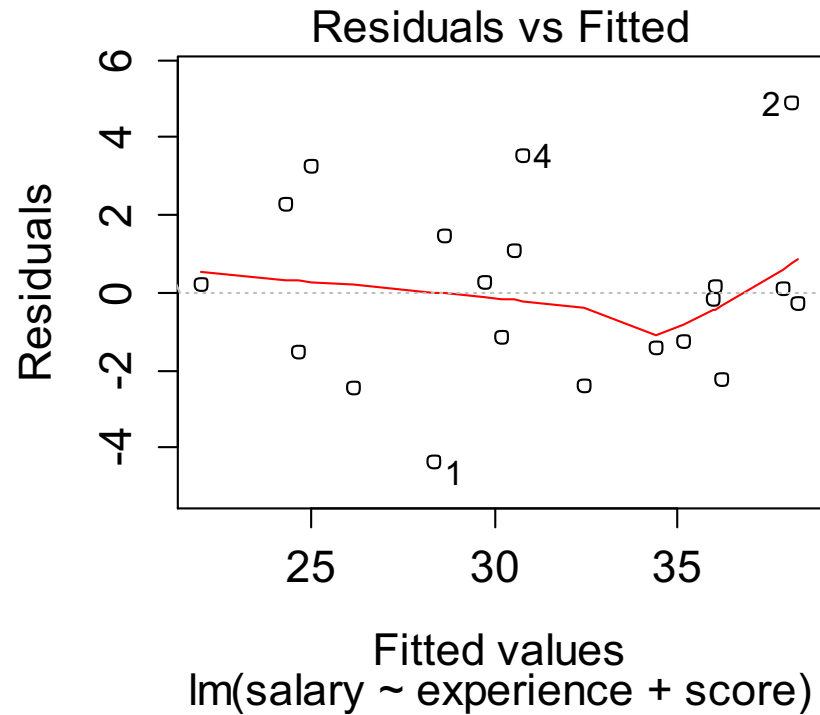
Multiple R-squared: 0.8342, Adjusted R-squared: 0.8147

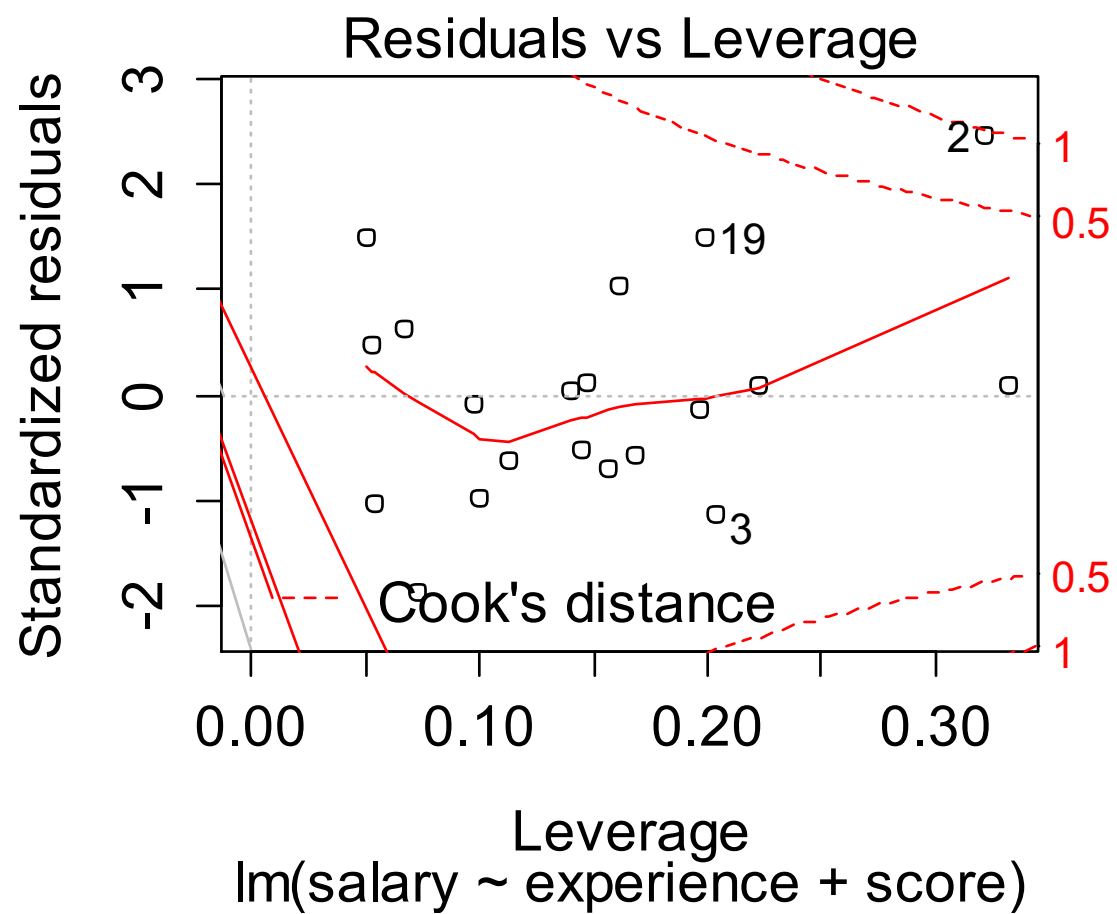
F-statistic: 42.76 on 2 and 17 DF, p-value: 2.328e-07

오차항에 대한 가정

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

잔차분석





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
----------	----------

다중공선성(multicollinearity)

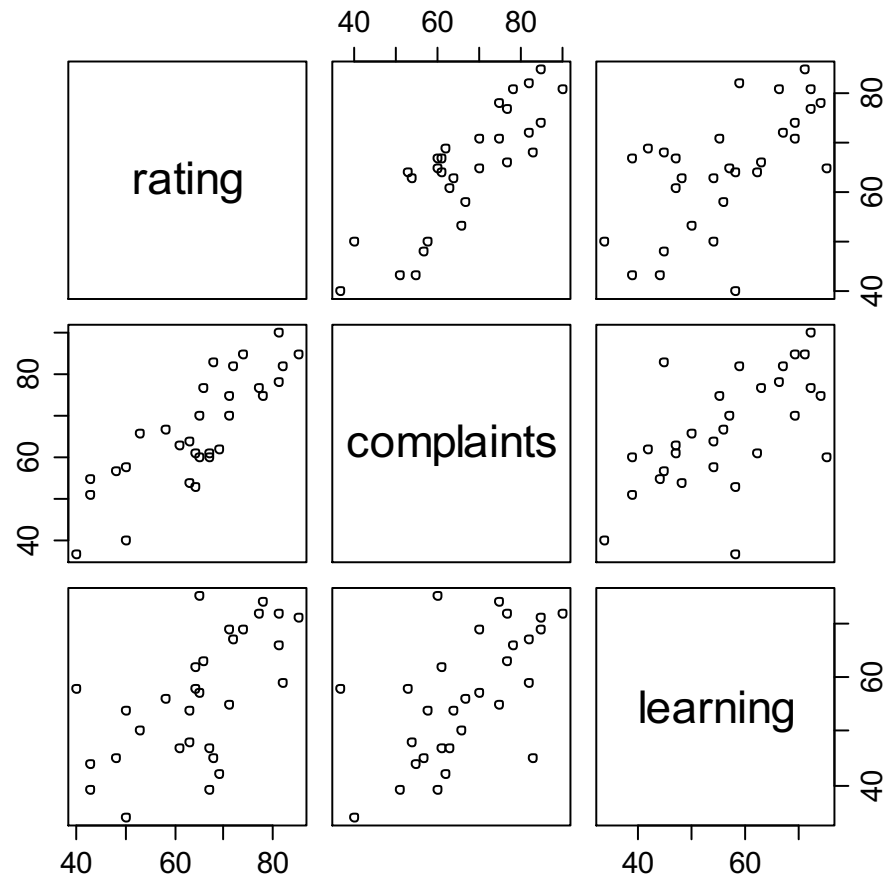
다중공선성은 독립변수들 사이의 상관관계를 지칭한다.

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

- Attitude data

한 금융회사의 30개 부서 직원들로부터 7 개의 사항에 대해 긍정적으로 대답한 비율을 포함 한 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]	advance	numeric	Advancement



■ 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

```
> 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 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, Mallows's C_p 혹은 최대 adjusted R^2 를 갖는 모형을 선택

모형선택 방법: 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

- 가장 유의하지 않은 변수 하나 제거 후 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 Selection의 자동화

```
> 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

모형선택 방법: 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	

모형선택 방법: 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

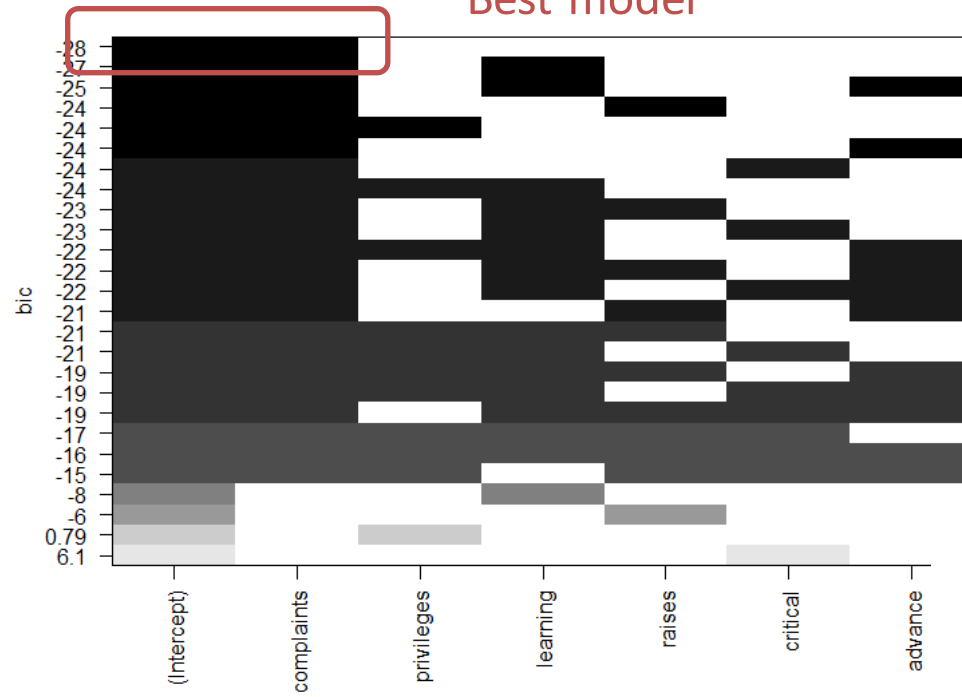
```
> summary(leaps)
Subset selection object
Call: regsubsets.formula(rating ~ ., data = attitude, nbest = 5)
6 variables (and intercept)
Forced in Forced out
complaints FALSE FALSE
privileges FALSE FALSE
learning FALSE FALSE
raises FALSE FALSE
critical FALSE FALSE
advance FALSE FALSE
5 subsets of each 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)	"*"	"*"	"*"	"*"	"*"	"*"

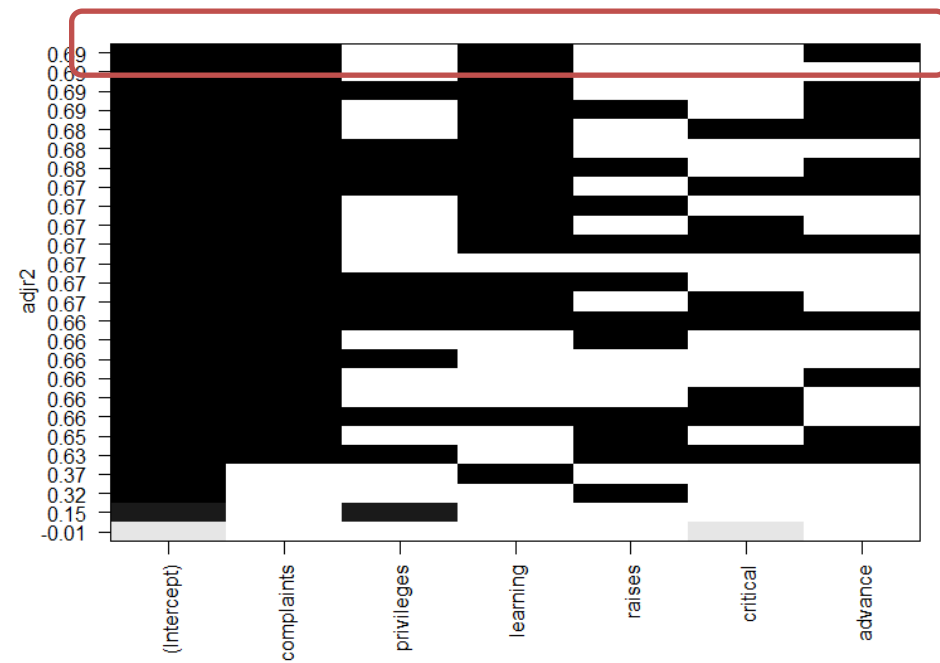
저장된 각 모형에
포함된 설명변수 표시

```
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
```

부적절한 다중회귀모형 활용

- 비선형 관계의 분석
- 잘못된 인과관계
 - 1인당 연간 국민소득으로 연간 자폐증 발생률을 설명
- 역인과관계
 - 골프 레슨을 받을 수록 나쁜 점수를 받음
- 변수 누락 편향
 - 골프 치는 사람들, 심장병, 암, 류머티즘 확률 높아
- 서로 관련이 깊은 설명변수 (다중공선성)
 - 불법약물 복용(헤로인, 코카인 복용)이 SAT 점수에 미치는 영향 (헤로인, 코카인 둘 중 하나만 하는 사람이 너무 적음)
- 데이터 범위를 벗어난 추정
 - 성인을 대상으로 신체활동 정도, 인종 등 변수를 활용해 체중 추정.
갓난아기의 체중 예측?