

R 기반 의학통계 및 머신러닝

박 승



충북대학교
CHUNGBUK NATIONAL UNIVERSITY

CHAPTER

03

임상연구에서의 회귀분석 활용

■ 회귀분석

- 회귀분석이란?

독립변수(설명 변수, 예측 변수)가 종속변수(반응 변수)에 미치는 영향을 분석하는 기법
특히 임상 연구에서는 변수 간의 관계를 수치화하고 인과관계를 추론하는 데 핵심적인 방법

- 변수 간의 관계 수치화 즉, 상관분석을 통해 변수간의 관계를 정립한 뒤, 논리적인 추론을 통해 두 변수 간 인과관계를 설명할 수 있음
- 독립변수의 편차가 커도 종속변수는 결국은 평균 근처로 회귀(Regression)한다는 뜻에서 ‘회귀분석’이라는 용어가 유래됨
- 현대 통계학에서는 수많은 독립변수들의 영향을 보정하고 변수간 관계를 규명하는데 있어서 널리 쓰임

■ 선형 회귀분석

▪ 단순 선형 회귀분석의 기본 모형

반응변수를 Y , 설명변수를 X 라 하면 다음과 같이 정의

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

또는 X, Y 에 대하여 관측된 n 쌍의 데이터 $(X_i, Y_i), i = 1, 2, \dots, n$ 를 사용하여 다음과 같이 표현

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

■ 선형 회귀분석

■ 선형 회귀분석의 기본 가정

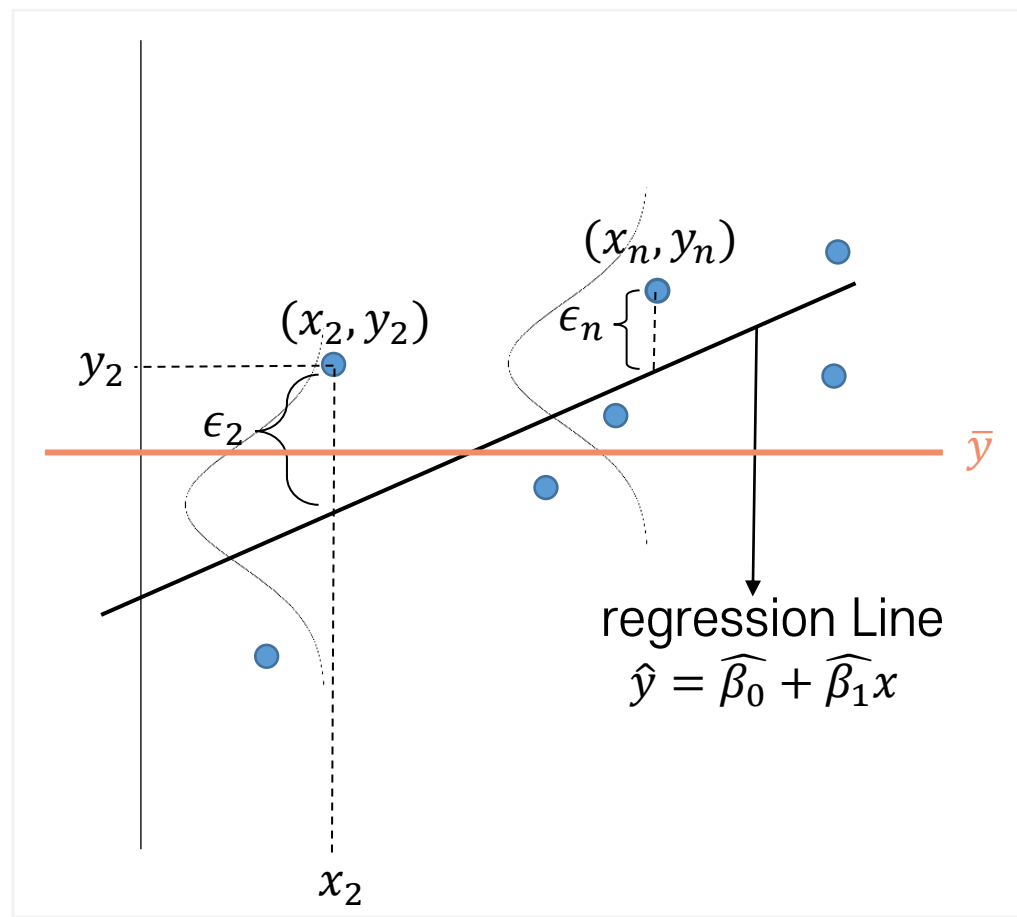
(1) 선형성: x 와 y 의 관계는 선형식으로 표현 가능

(2) 독립성: 오차 ϵ_i 는 서로 독립

(3) 정규성: 오차 ϵ_i 는 정규분포를 따름

(4) 등분산성: 오차 ϵ_i 는 동일한 분산을 가짐

$$\Rightarrow \epsilon_i \sim iid N(0, \sigma^2), \quad i = 1, 2, \dots, n$$



■ 선형 회귀분석

■ 단순 선형 회귀분석의 계수 추정

회귀계수: y -절편 β_0 와 회귀선의 기울기 β_1

관측된 데이터를 이용하여 회귀계수 β_0 , β_1 의 추정과 검정

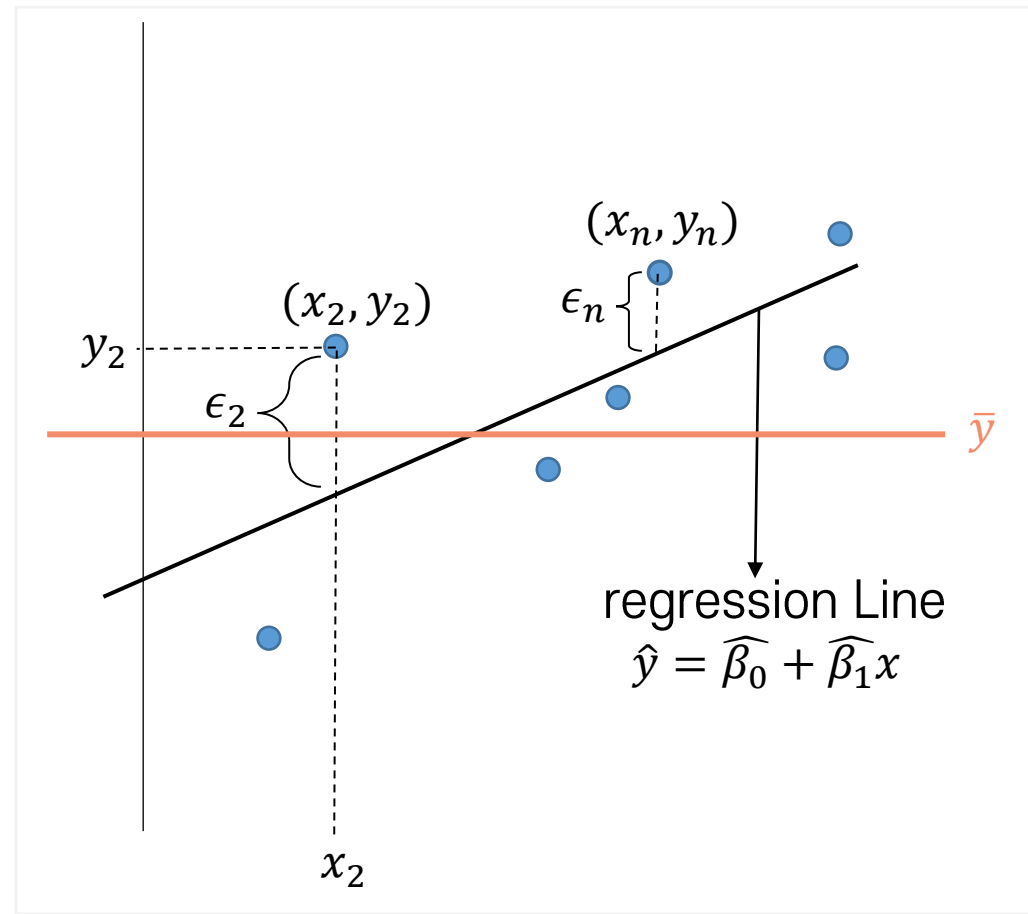
기울기 β_1 은 x 가 한 단위 증가할 때 y 가 얼마나 증가 또는 감소하는 지를 나타내므로
일반적으로 β_0 보다 β_1 에 관심을 더 갖게 됨

■ 선형 회귀분석

■ 최소제곱법 (Method of Least Square)

- 회귀선의 오차 ϵ 를 최소화 하는 것이 목표
- ϵ 는 음수 양수 모두 존재하기 때문에
제곱하여 미분한 뒤 최소가 되는 $\hat{\beta}_0, \hat{\beta}_1$ 을 추정

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$



■ 선형 회귀분석

$$Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{dQ}{d\beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) = 0, \quad \frac{dQ}{d\beta_1} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i) X_i = 0$$

정규방정식: 위 식을 만족하는 β_0, β_1 의 값을 $\widehat{\beta}_0, \widehat{\beta}_1$ 으로 대체한 뒤 $\widehat{\beta}_0, \widehat{\beta}_1$ 를 구함

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

****다중 회귀 분석의 경우 벡터에 같은 방식을 적용**

03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

▪ 예제

- 13명의 아버지와 이들의 아들의 키(단위:cm)를 측정한 자료가 다음과 같다.

단, 사람의 키는 정규분포를 따른다고 가정한다.

아버지의 키를 x , 아들의 키를 y 라 할 때, 단순 선형회귀모형을 구하시오.

	1	2	3	4	5	6	7	8	9	10	11	12	13
아버지	168	160	170	158	176	161	180	183	180	167	179	171	166
아들	179	169	180	160	178	170	183	187	179	172	181	173	165

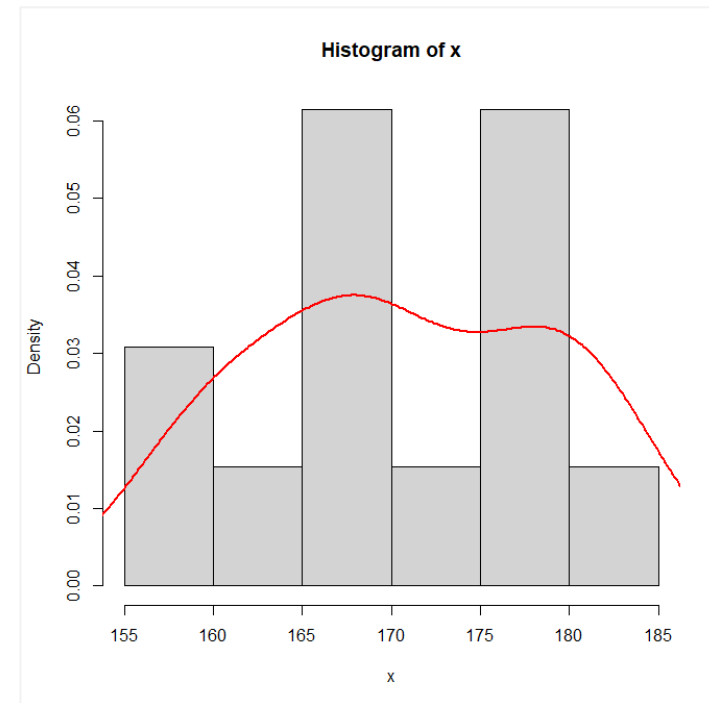
03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

- 데이터 탐색
- 히스토그램(막대그래프)

```
x<-c(168,160,170,158,176,161,180,183,180,167,179,171,166)
y<-c(179,169,180,160,178,170,183,187,179,172,181,173,165)

hist(x, freq=F)
lines(density(x), col="red", lwd=2) |
```

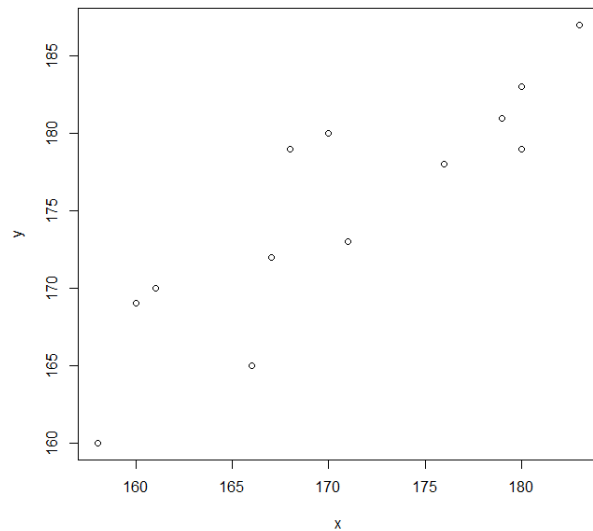


03 임상연구에서의 회귀분석 활용

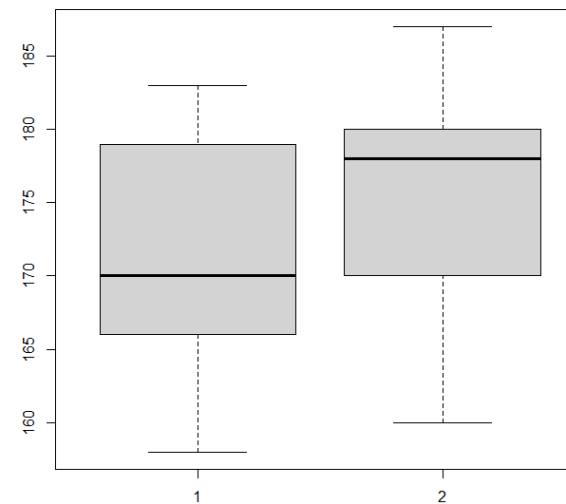
■ 선형 회귀분석

■ 데이터 탐색

산점도 `plot(x, y)`



상자그림 `boxplot(x, y)`



■ 선형 회귀분석

```
x<-c(168,160,170,158,176,161,180,183,180,167,179,171,166)
y<-c(179,169,180,160,178,170,183,187,179,172,181,173,165)
```

#X,Y 자료 입력
회귀분석은 서로 대응되는
쌍의 자료로 이루어지므로
순서가 바뀌면 안됨

```
length(x)
length(y)    #X,Y 자료의 길이 비교
```

```
lm(y~x)    #Y를 종속변수로, X를 독립변수로 하는 선형 회귀 모형 적합
summary(lm(y~x))    #해당 모형의 내용 요약
```

■ 선형 회귀분석

다른 방법(데이터프레임 이용)

```
x<-c(168,160,170,158,176,161,180,183,180,167,179,171,166) #X,Y 자료 입력
y<-c(179,169,180,160,178,170,183,187,179,172,181,173,165)
```

```
ex1<- cbind(x,y) #X,Y 자료 합치기
str(ex1)
```

```
ex1<- data.frame(ex1) #데이터 프레임 형태로 변환, 생략한 경우 오류 발생
```

```
modell<-lm(y~x, (data=ex1)) #ex1 데이터 내부에 있는 y,x를 불러오는 경우 data= 항목이 필요
summary(modell)
```

```
> modell<-lm(y~x, data=ex1)
```

```
model.frame.default(formula = y ~ x, data = ex1, drop.unused.levels = TRUE)에서 다음과 같은 에러가 발생했습니다:
'data'는 행렬 또는 배열이 아닌 반드시 데이터 프레임이어야 합니다
```

03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

결과 해석

```
> model1
```

```
Call:
```

```
lm(formula = y ~ x, data = ex1) #  $Y = \beta_0 + \beta_1 X$  모델 형태를 의미
```

```
Coefficients:
```

```
(Intercept)  
37.8090
```

```
x  
0.8042
```

절편의 값
즉 β_0 를 의미함

x 의 계수
즉 β_1 값을 의미함



아들의 키 = $37.809 + 0.804 \times$ 아버지의 키

■ 선형 회귀분석

```
> summary(model1)
```

```
Call:
```

```
lm(formula = y ~ x, data = ex1)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max    #잔차의 분포 요약
-6.3034 -2.3244 -0.1076  2.5217  6.0882
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.8090    23.1870    1.631    0.131
x             0.8042     0.1357    5.927 9.92e-05 ***
---

```

#계수의 유의성 검정

β_1 의 검정의 경우, 귀무가설은 $H_0 : \beta_1 = 0$

t value의 경우, 실제 x의 값에 따라 도출된 검정통계량
귀무가설이 참이라는 전제하에 해당 검정통계량이 나올
확률이 0.0000992

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.92 on 11 degrees of freedom
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.7398
F-statistic: 35.12 on 1 and 11 DF,  p-value: 9.923e-05
```

#분산분석표와 함께 설명

03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

```
> summary(aov(model1))
              Df Sum Sq Mean Sq F value    Pr(>F)    
x               1   539.9    539.9    35.12 9.92e-05 ***
Residuals     11   169.1     15.4              
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

비고	제곱합	자유도	평균제곱합	F-값
회귀제곱합	539.86	1	539.86	35.12635
오차제곱합	169.06	11	15.36909	
총제곱합	708.92	12		

$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$, 회귀직선으로 설명되는 부분

$\sum_{i=1}^n (Y_i - \bar{Y})^2$,
총 변동량

샘플 수(13) - 1

추정한 계수의 개수(β_0, β_1 2개) - 1

--> #회귀모형의 적합도 검정

03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

비고	제곱합	자유도	평균제곱합	F-값
회귀제곱합	539.86	1 ^①	539.86	35.12635
오차제곱합	169.06	11	^⑤ 15.36909	
총제곱합	708.92	12		

Residual standard error: ^② 3.92 on 11 degrees of freedom
 Multiple R-squared: ^③ 0.7615, Adjusted R-squared: 0.7398
 F-statistic: ^④ 35.12 on 1 and 11 DF, p-value: 9.923e-05

^③ $R^2 = \frac{SS_{Reg}}{SS_{Total}}$, 총 제곱합 중 회귀제곱합의 비율
 단순선형 회귀의 경우, 상관계수(ρ)의 제곱 = R^2

```
> cor(x, y)
[1] 0.8726475
```

^② #Residual standard error는 평균제곱합의 제곱근

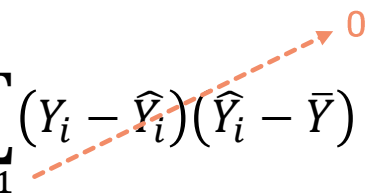
#F-통계량 값은 35.12635로, $F_{1,11}$ 분포를 따름. 단순선형 회귀의 경우 t-값의 제곱이 F-값. ($5.927^2 \approx 35.129$)

$$\Rightarrow t_r^2 \sim \left(\frac{Z}{\sqrt{\frac{W}{r}}} \right)^2 = \frac{Z^2/1}{W/r} \sim F_{1,r}$$

다시 말해, β_1 의 유의성 검정(회귀계수에 대한 검정)이 모형의 적합성 검정과 같은 의미

■ 선형 회귀분석

총 제곱합 = 잔차 제곱합 + 회귀 제곱합

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$


$$\therefore \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

SST (총제곱합) = SSE (잔차제곱합) + SSReg (회귀제곱합)

** 데이터의 총 변동량 즉, 총 제곱합은 회귀제곱합과 잔차제곱합의 두 부분으로 나눌 수 있음

회귀선으로 설명되는 변동량 = 회귀제곱합, 오차로 설명되는 변동량 = 잔차제곱합

■ 선형 회귀분석

총 제곱합 = 잔차 제곱합 + 회귀 제곱합

- R^2 와 제곱합 간의 관계

R^2 값은 총 변동량 중 회귀제곱합의 비율

$$R^2 = \frac{\text{회귀 제곱합}}{\text{총 제곱합}} = \frac{SS_{Reg}}{SS_{Total}}$$

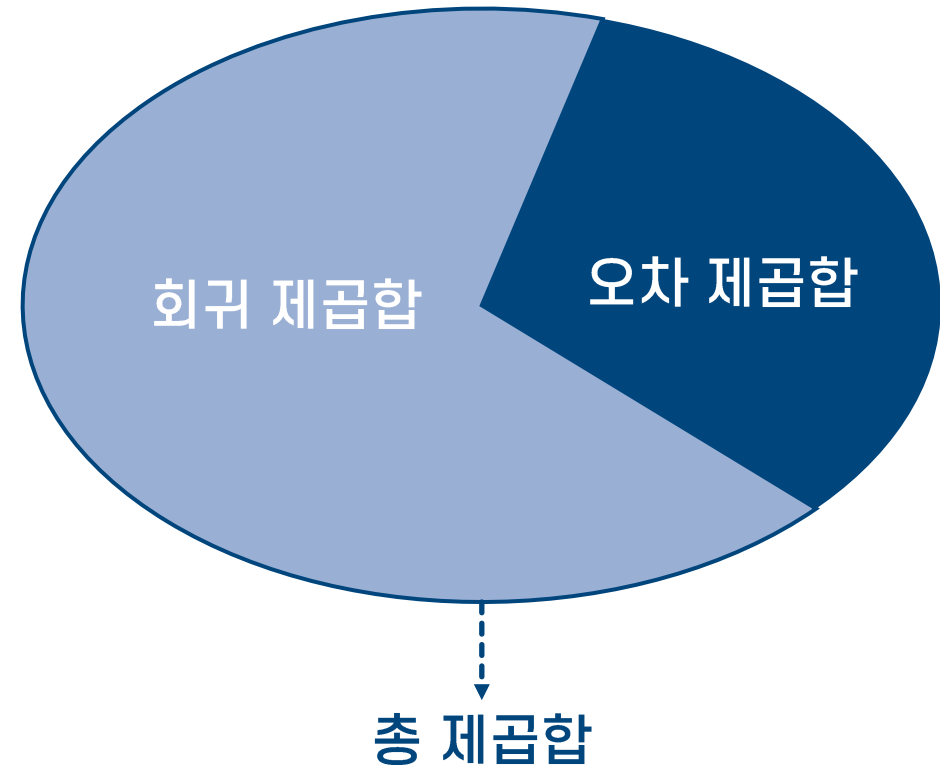
즉, R^2 값은 회귀선으로 표현 가능한 데이터의 비율

- 데이터가 회귀선으로 설명되는 비중이 클 수록

R^2 값이 높아짐

- 반대로, 데이터가 회귀선으로 잘 설명되지 않을 경우

즉, 오차 제곱합이 클 수록 R^2 값은 낮아짐



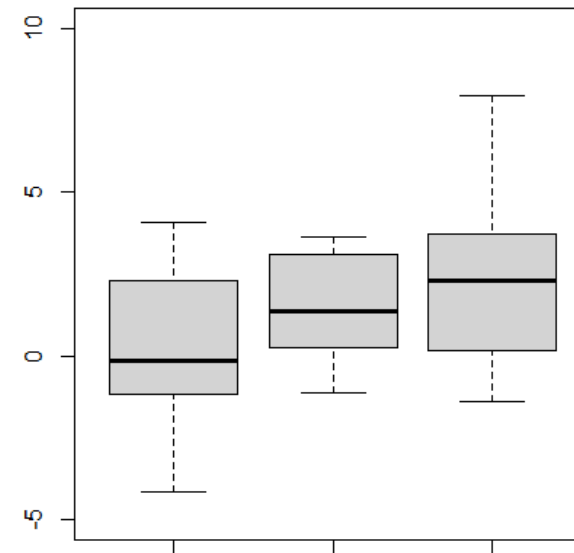
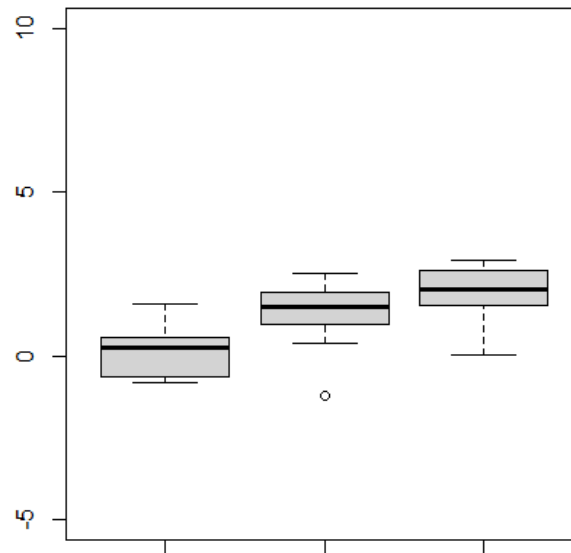
03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

총 제공합 = 잔차 제공합 + 회귀 제공합

▪ example)

- 아래와 같은 상자 그림을 살펴 보자. 각각 모집단의 평균은 각각 0, 1, 2으로 동일하며, 분산만 다르다. 세 집단의 평균이 같지 않다는 귀무가설을 설정했을 때, 이를 기각할 확률은 어느 쪽이 높을까?



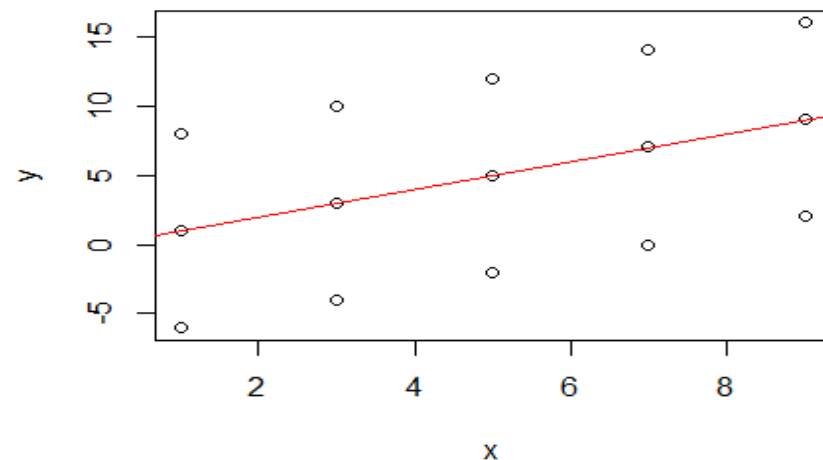
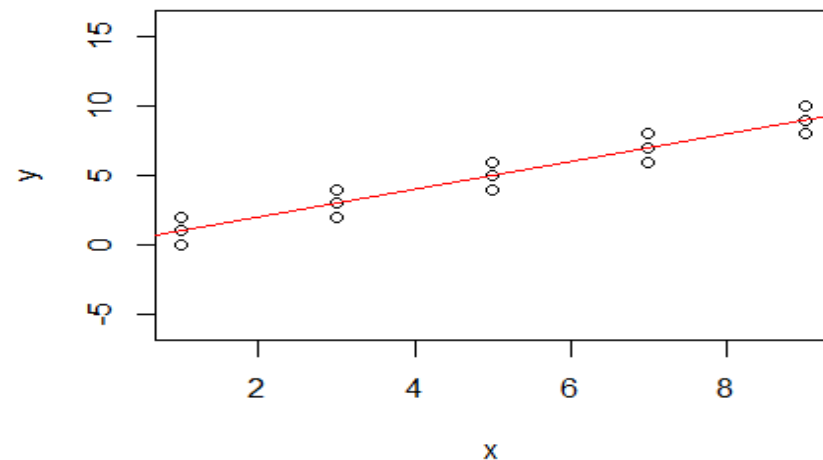
03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

총 제곱합 = 잔차 제곱합 + 회귀 제곱합

■ example)

- 오른쪽의 두 산점도+회귀직선 그림을 비교해보자
- 두 그룹의 회귀직선은 동일하다
- 두 그룹의 절편도 동일하다
- 이때 회귀 직선의 계수 즉, β_1 가 0이다(= 회귀 직선이 유의하지 않다) 라는 귀무가설을 설정했을 때, 귀무가설을 기각할 가능성이 높은 것은 위와 아래 중 어느쪽일까?



■ 선형 회귀분석

총 제공합 = 잔차 제공합 + 회귀 제공합

■ 실습)

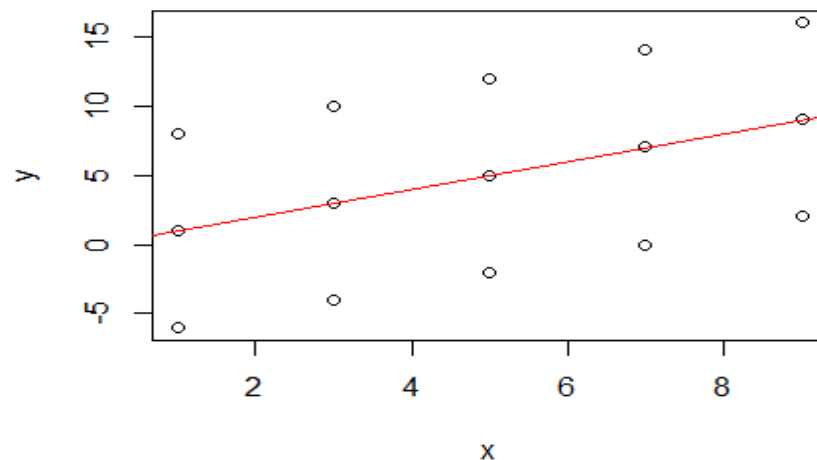
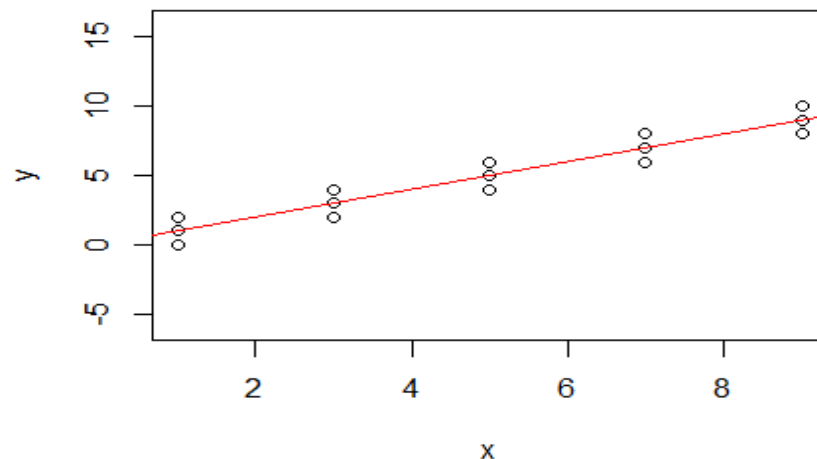
```
library(dplyr)
par(mfrow=c(2,1))

x<-c(1,1,1,3,3,3,5,5,5,7,7,7,9,9,9)
y<-c(0,1,2,2,3,4,4,5,6,6,7,8,8,9,10)

plot(x,y, ylim=c(-6,16))
abline(lm(y~x), col="red")
lm(y~x) %>% summary
lm(y~x) %>% aov %>% summary

u<-c(1,1,1,3,3,3,5,5,5,7,7,7,9,9,9)
v<-c(-6,1,8,-4,3,10,-2,5,12,0,7,14,2,9,16)

plot(u,v,ylim=c(-6,16))
abline(lm(v~u), col="red")
lm(v~u) %>% summary
lm(v~u) %>% aov %>% summary
```



03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

총 제공합 = 잔차 제공합 + 회귀 제공합

■ 실습)

- 왼쪽 그룹의 R^2 값은 0.917, 오른쪽 그룹의 R^2 값은 0.135로 나타남
- 두 그룹은 동일한 회귀직선을 가지고 있음에도 불구하고, 오차항의 크기 차이 때문에 다른 결과가 나타남

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.17e-16   4.60e-01    0.0      1
x             1.00e+00   8.01e-02   12.5   1.3e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.877 on 13 degrees of freedom
Multiple R-squared:  0.923,    Adjusted R-squared:  0.917
F-statistic: 156 on 1 and 13 DF, p-value: 1.29e-08
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	120	120.0	156	1.3e-08 ***
Residuals	13	10	0.8		

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83e-15   3.22e+00    0.00    1.000
u             1.00e+00   5.60e-01    1.78    0.098 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.14 on 13 degrees of freedom
Multiple R-squared:  0.197,    Adjusted R-squared:  0.135
F-statistic: 3.18 on 1 and 13 DF, p-value: 0.0977
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
u	1	120	120.0	3.18	0.098 .
Residuals	13	490	37.7		

03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

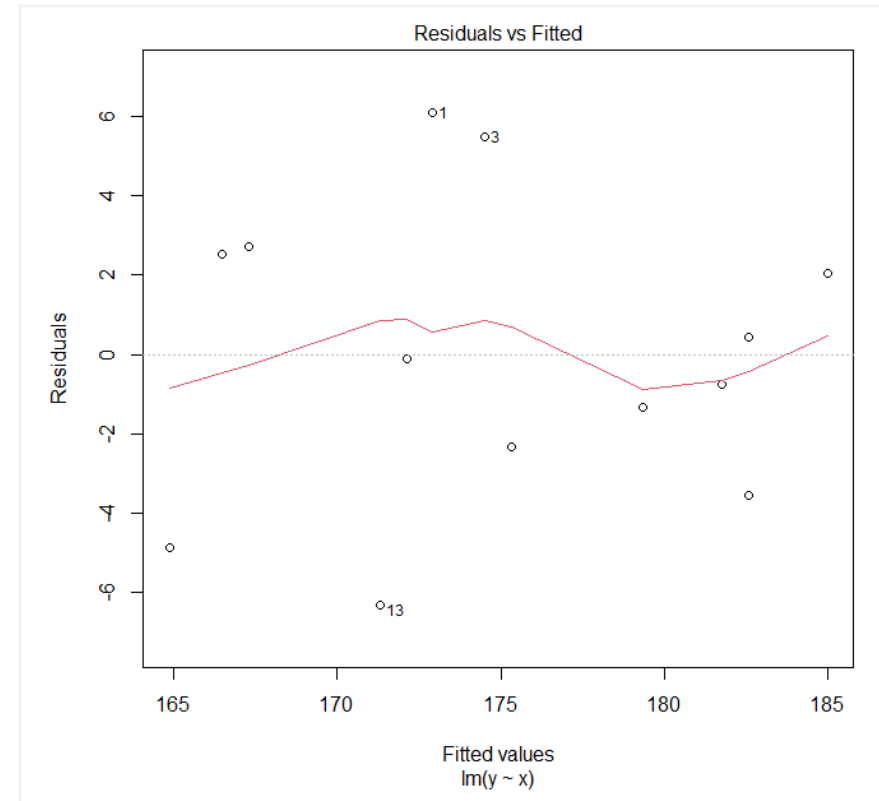
- 잔차분석 `plot(model1)`
- residuals vs. fitted values

잔차와 적합값의 비교

잔차의 독립성 및 추세를 확인 가능

잔차가 특정한 패턴을 보이는 경우 독립성 위배

- 시계열 분석 등의 다른 분석 방법 도입 필요



03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

- 잔차분석 `plot(model1)`
- Q-Q plot

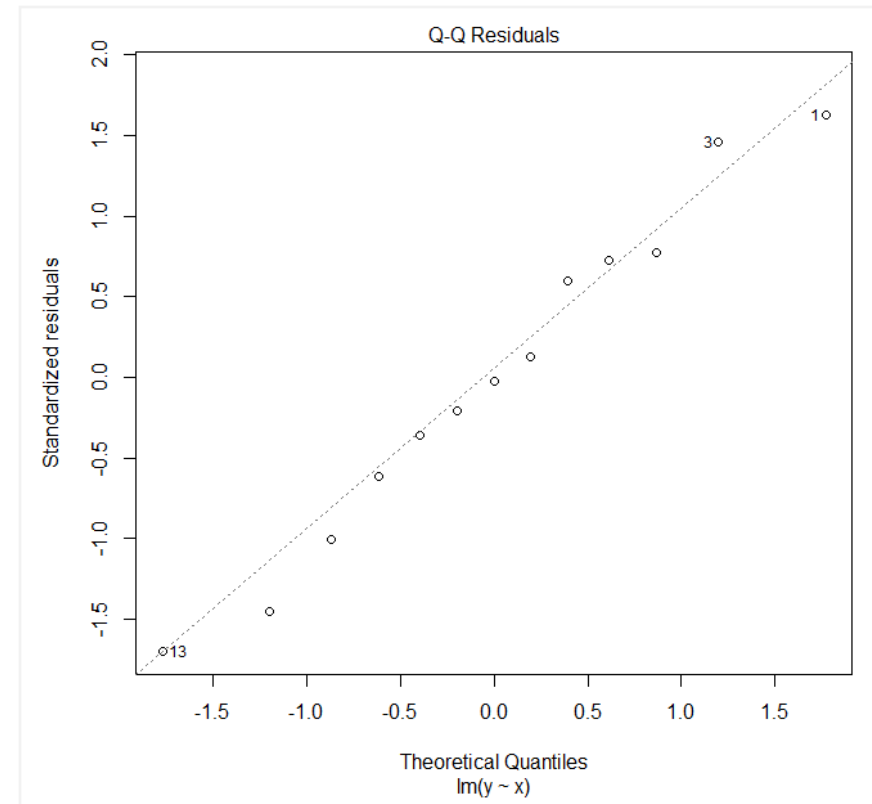
잔차의 정규성 검정

`shapiro.test(model1$residuals)`

대각선에서 멀리 위치하는 경우

잔차의 정규성 가정이 위배될 수 있음

- 변수 변환등의 방법을 통해 정규성 확보 필요



03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

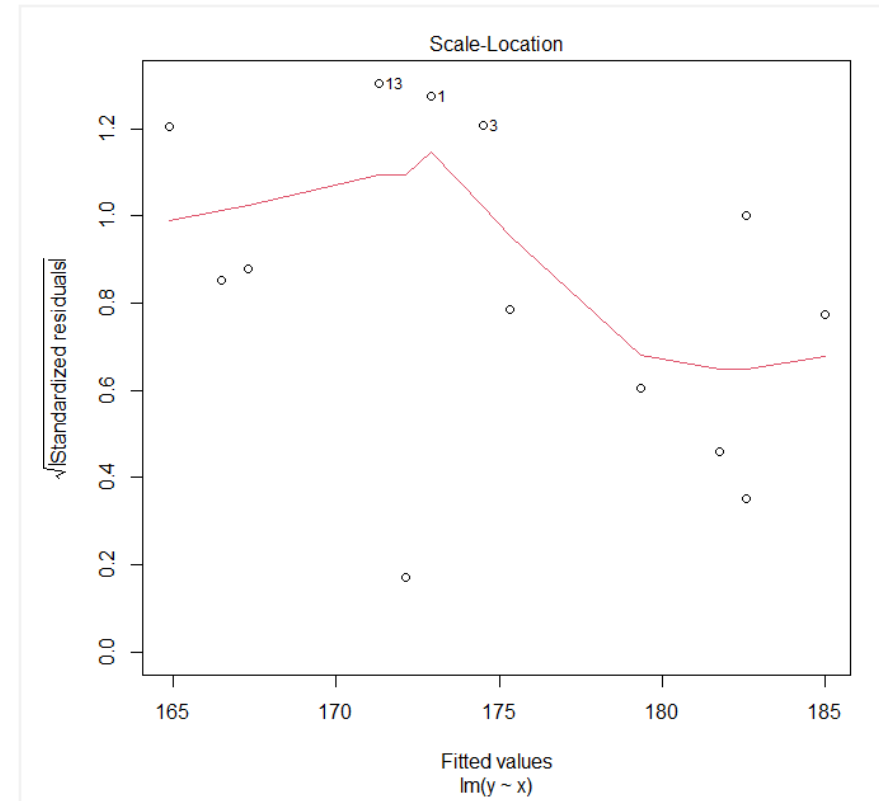
- 잔차분석 plot(model1)
- scale-Location

잔차의 등분산성 검정

적합값(예측값)에 따른 잔차의 변화를 보여줌

붉은 선이 추세를 띠거나 요동치면 등분산성 위배

- 변수 변환등의 방법을 통해 등분산성 확보 필요



03 임상연구에서의 회귀분석 활용

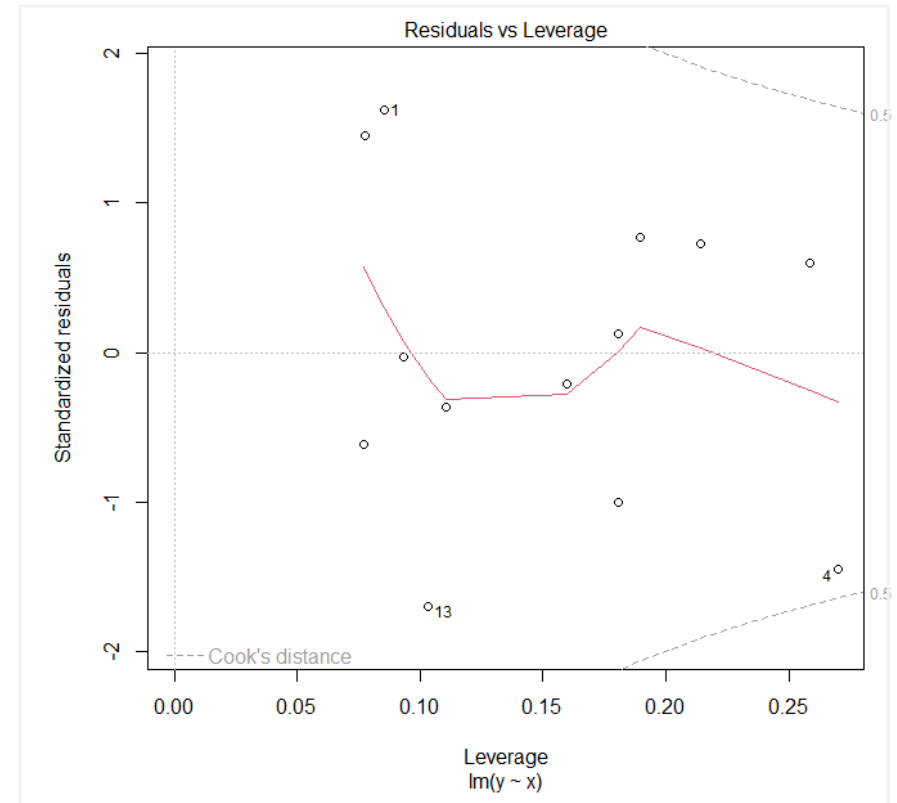
■ 선형 회귀분석

- 잔차분석 plot(model1)
- Residuals vs. Leverage

이상치 또는 극단값 확인

Cook's Distance 라인 밖에 위치하면 극단값으로 간주

- 해당 샘플 확인 후 제거 또는 보정 필요



03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

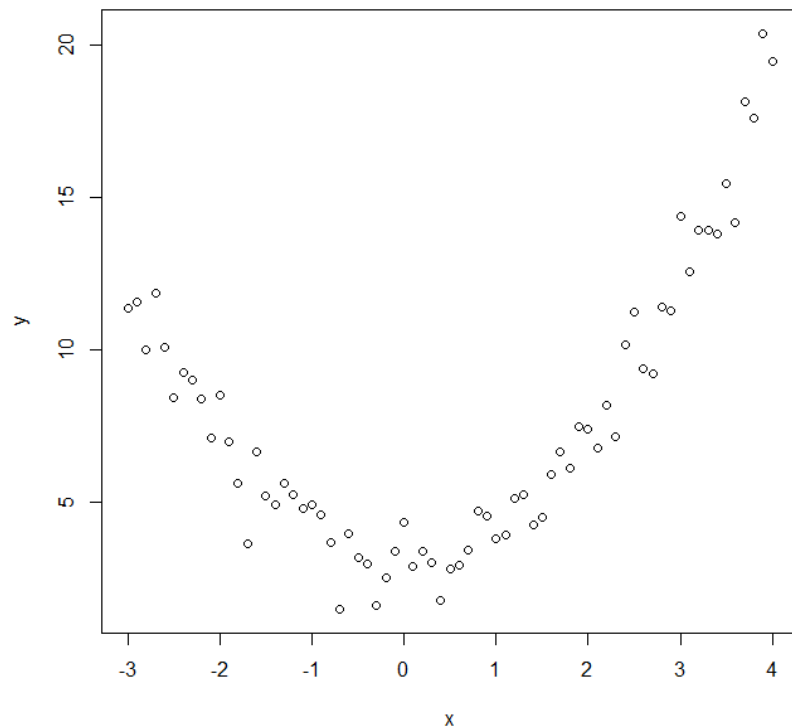
- 변수변환
- 다음과 같은 자료를 고려하는 경우

```
set.seed(1) #시드 고정
```

```
x<- seq(-3,4, by=0.1) #공차 0.1인 수열 생성
```

```
y<- 3+x^2+rnorm(length(x),0,1)
```

```
plot(x,y) #정규분포 랜덤 샘플 생성
```



03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

- 변수변환
- 제곱형태의 자료

Residuals:

Min	1Q	Median	3Q	Max
-5.724	-3.290	-1.133	3.359	9.238

#잔차의 차이가 큼

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.0915	0.4901	14.471	< 2e-16 ***
x	1.0394	0.2323	4.474	2.95e-05 ***

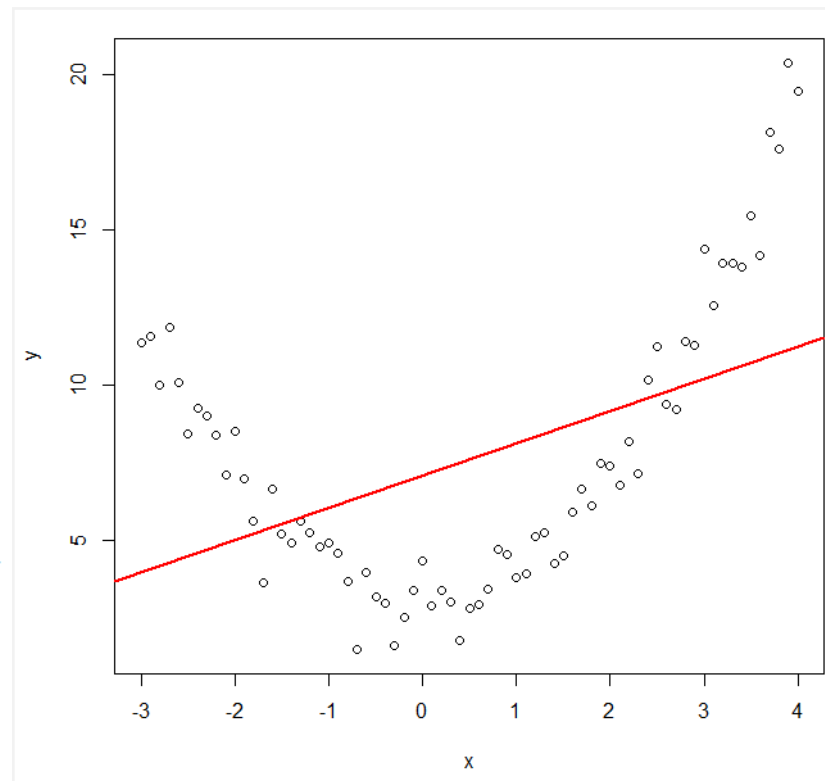
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.012 on 69 degrees of freedom

Multiple R-squared: 0.2249, Adjusted R-squared: 0.2136

F-statistic: 20.02 on 1 and 69 DF, p-value: 2.946e-05

R^2 값이 0.22로 저조

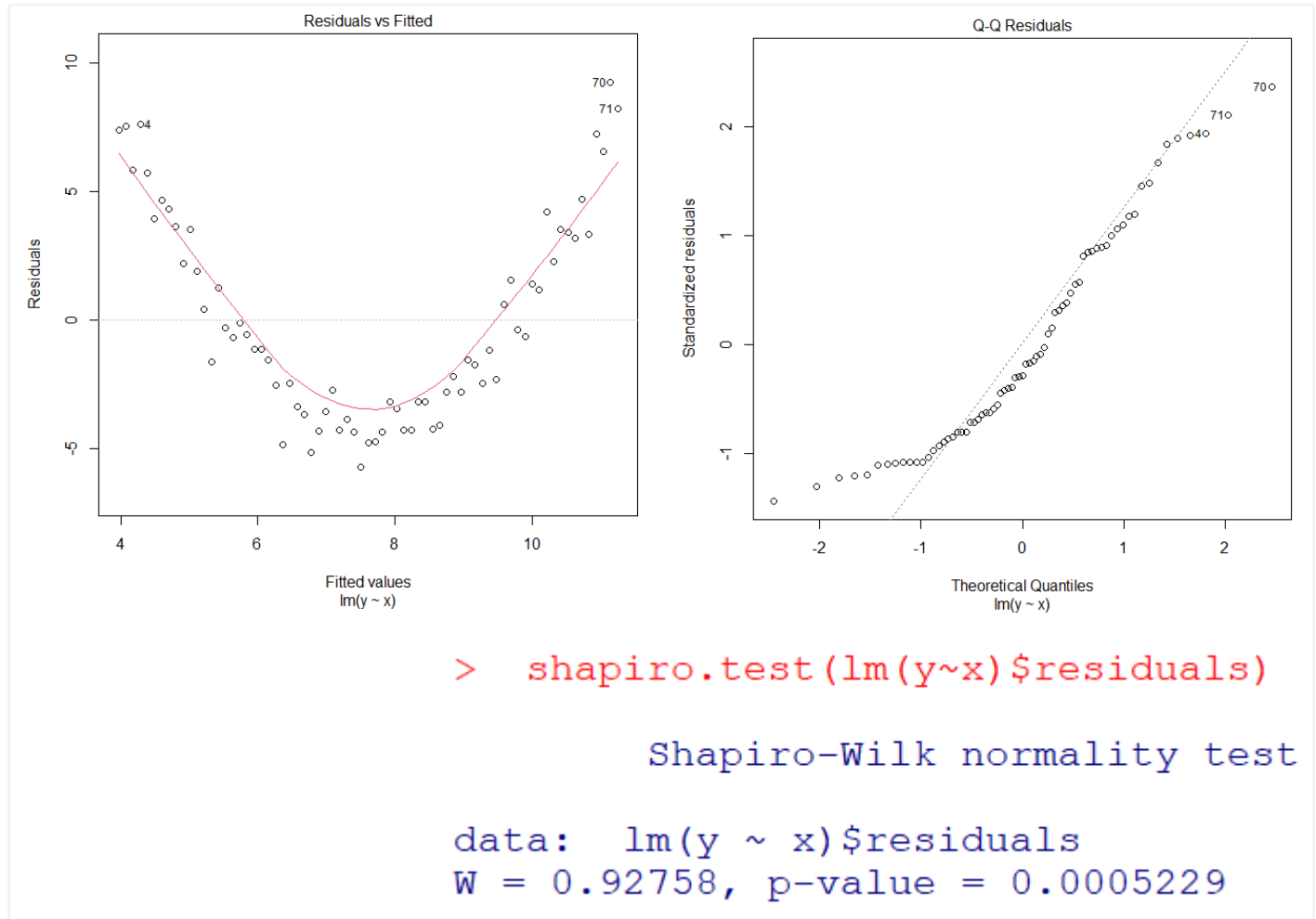


03 임상연구에서의 회귀분석 활용

■ 선형 회귀분석

- 변수변환
- 제곱형태의 자료 잔차분석

잔차분석 결과
잔차의 추세가 확인되며
정규성도 위배됨



■ 선형 회귀분석

- 변수변환
- 제곱형태의 모형 수립

$$Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i$$

```
summary(lm(y~I(x^2)))
```

y~x^2로 타이핑할 경우,
^2를 상호작용으로 인식하기때문에 I(x^2)로 표시해주어야 함

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.33160 -0.46119 -0.08998  0.59800  2.11124

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.03485     0.15721   19.30  <2e-16 ***
I(x^2)        1.02839     0.02547   40.38  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9181 on 69 degrees of freedom
Multiple R-squared:  0.9594,    Adjusted R-squared:  0.9588
F-statistic: 1631 on 1 and 69 DF,  p-value: < 2.2e-16
```

- 변수변환
- 제곱형태의 모형 수립

The figure displays two diagnostic plots for a linear regression model.

The left plot is a scatter plot of the response variable y against the predictor variable x^2 . The data points are represented by open circles, and a solid blue line indicates the fitted regression line. The x-axis ranges from 0 to 15, and the y-axis ranges from 0 to 20.

The right plot is a residual plot titled "Residuals vs Fitted". The residuals are plotted against the fitted values, with a red smoothing line overlaid. The x-axis is labeled "Fitted values" and "lm(y ~ l(x^2))", ranging from 0 to 20. The y-axis is labeled "Residuals", ranging from -2 to 2. Three points are specifically labeled: 14, 61, and 67.

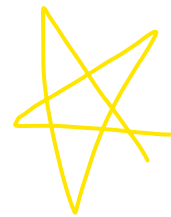
- ```
> shapiro.test(lm(y~I(x^2))$residuals)
```

Shapiro-Wilk normality test

```
data: lm(y ~ I(x^2))$residuals
W = 0.9891, p-value = 0.8003
```



### ■ 선형 회귀분석\_회귀분석에서의 선형성



#### 1. 변수변환을 통해 선형이 되는 경우는 선형으로 간주

$$\text{ex) } y = \beta_1 \exp(x) \quad \exp(x) = t \Rightarrow y = \beta_1 t$$

$$y = 1 - \exp\left(-\left(\frac{x}{\eta}\right)^\beta\right) \quad \ln x = t \Rightarrow \ln\{-\ln(1 - y)\} = -\beta \ln \eta + \beta t$$

#### 2. 변수변환을 해도 선형이 되지 않는 경우는 비선형

$$\text{ex) } y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (\text{로지스틱 모형})$$

### ■ 다중 회귀분석

- 데이터 로드
- 내장데이터 이용(mtcars)
- ?mtcars 를 통해 데이터에 관한  
변수 설명 등의 도움말을 얻을 수 있음  
(해당 도움말이 작동하지 않는 경우  
R을 관리자 권한으로 실행 시도)

#### Format

A data frame with 32 observations on 11 (numeric) variables.

```
[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 Engine (0 = V-shaped, 1 = straight)
[9] am Transmission (0 = automatic, 1 = manual)
[10] gear Number of forward gears
[11] carb Number of carburetors
```

### ■ 다중 회귀분석

- 데이터 탐색
- str(mtcars)를 통해  
데이터의 형태를 파악 가능
  - 맞춤 EDA 전략 수립

```
> str(mtcars)
'data.frame': 32 obs. of 11 variables:
 $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num 160 160 108 258 360 ...
 $ hp : num 110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt : num 2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num 16.5 17 18.6 19.4 17 ...
 $ vs : num 0 0 1 1 0 1 0 1 1 1 ...
 $ am : num 1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```

## 03 임상연구에서의 회귀분석 활용

### ■ 다중 회귀분석

#### ▪ 데이터 탐색

- 독립변수간 상관계수가 대체로 높게 나타남
- 절대값 0.5~0.7의 상관계수가 많음

```
library(writexl)
write_xlsx(data.frame(cor(mtcars)), "0228.xlsx")
```

|      | mpg   | cyl   | disp  | hp    | drat  | wt    | qsec  | vs    | am    | gear  | carb  |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| mpg  | 1     | -0.85 | -0.85 | -0.78 | 0.68  | -0.87 | 0.42  | 0.66  | 0.60  | 0.48  | -0.55 |
| cyl  | -0.85 | 1     | 0.90  | 0.83  | -0.70 | 0.78  | -0.59 | -0.81 | -0.52 | -0.49 | 0.53  |
| disp | -0.85 | 0.90  | 1     | 0.79  | -0.71 | 0.89  | -0.43 | -0.71 | -0.59 | -0.56 | 0.39  |
| hp   | -0.78 | 0.83  | 0.79  | 1     | -0.45 | 0.66  | -0.71 | -0.72 | -0.24 | -0.13 | 0.75  |
| drat | 0.68  | -0.70 | -0.71 | -0.45 | 1     | -0.71 | 0.09  | 0.44  | 0.71  | 0.70  | -0.09 |
| wt   | -0.87 | 0.78  | 0.89  | 0.66  | -0.71 | 1     | -0.17 | -0.55 | -0.69 | -0.58 | 0.43  |
| qsec | 0.42  | -0.59 | -0.43 | -0.71 | 0.09  | -0.17 | 1     | 0.74  | -0.23 | -0.21 | -0.66 |
| vs   | 0.66  | -0.81 | -0.71 | -0.72 | 0.44  | -0.55 | 0.74  | 1     | 0.17  | 0.21  | -0.57 |
| am   | 0.60  | -0.52 | -0.59 | -0.24 | 0.71  | -0.69 | -0.23 | 0.17  | 1     | 0.79  | 0.06  |
| gear | 0.48  | -0.49 | -0.56 | -0.13 | 0.70  | -0.58 | -0.21 | 0.21  | 0.79  | 1     | 0.27  |
| carb | -0.55 | 0.53  | 0.39  | 0.75  | -0.09 | 0.43  | -0.66 | -0.57 | 0.06  | 0.27  | 1     |

⇒ 다중공선성(multicollinearity): 독립변수들 간 강한 상관관계가 있는 경우 발생하는 문제

## ■ 다중 회귀분석

### ■ 모형 적합

```
model2<-lm(mpg ~. ,data=mtcars)
summary(model2)
```

### ■ $R^2$ 값은 0.869

$R^2_{adj}$  값은 0.807로 높지만

모든 변수가 유의하지 않게 나타남

### ■ 수정 결정계수: $R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$

### ■ 결정계수 $R^2$ 은 변수 개수가 늘어나면 항상 증가하는 단점이 있음

수정결정계수  $R^2_{adj}$ 는 독립변수 개수를 고려하여 패널티로 보정한 값

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t ) |
|-------------|----------|------------|---------|----------|
| (Intercept) | 12.30337 | 18.71788   | 0.657   | 0.5181   |
| cyl         | -0.11144 | 1.04502    | -0.107  | 0.9161   |
| disp        | 0.01334  | 0.01786    | 0.747   | 0.4635   |
| hp          | -0.02148 | 0.02177    | -0.987  | 0.3350   |
| drat        | 0.78711  | 1.63537    | 0.481   | 0.6353   |
| wt          | -3.71530 | 1.89441    | -1.961  | 0.0633   |
| qsec        | 0.82104  | 0.73084    | 1.123   | 0.2739   |
| vs          | 0.31776  | 2.10451    | 0.151   | 0.8814   |
| am          | 2.52023  | 2.05665    | 1.225   | 0.2340   |
| gear        | 0.65541  | 1.49326    | 0.439   | 0.6652   |
| carb        | -0.19942 | 0.82875    | -0.241  | 0.8122   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom

Multiple R-squared: 0.869, Adjusted R-squared: 0.8066

F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

## 03 임상연구에서의 회귀분석 활용

### ■ 다중 회귀분석

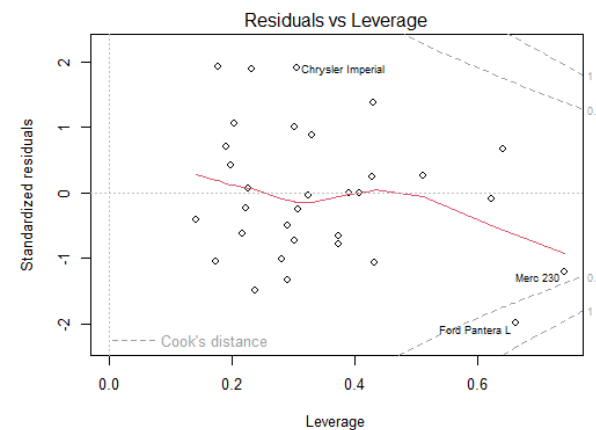
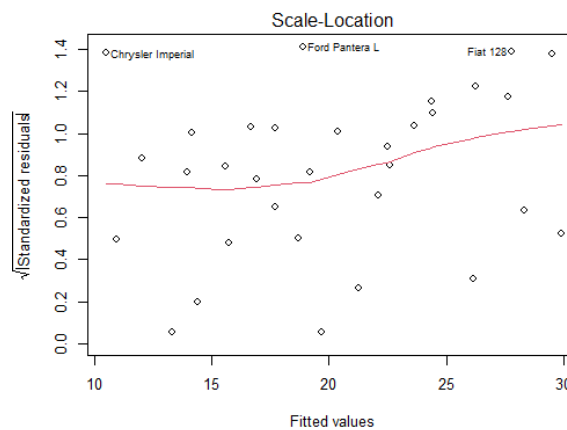
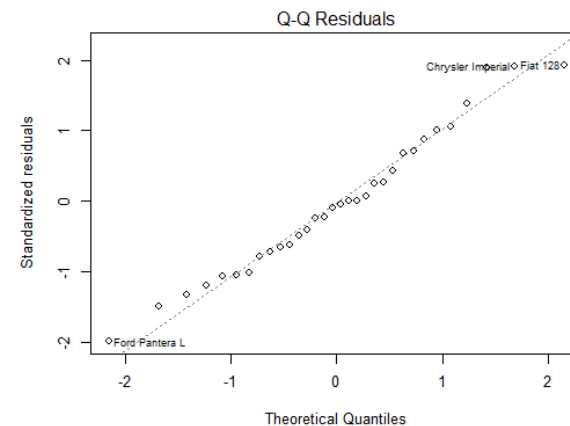
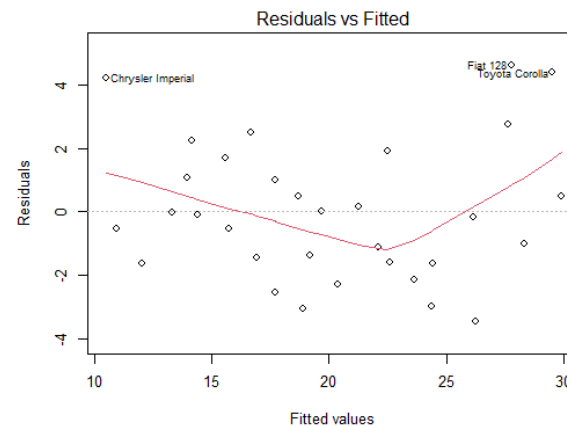
#### ■ 모형 진단(잔차분석)

```
par(mfrow=c(2,2)) #plot을 2by2로 출력
plot(model2)
```

```
> shapiro.test(model2$residuals)
```

Shapiro-Wilk normality test

```
data: model2$residuals
W = 0.95694, p-value = 0.2261
```



### ■ 다중 회귀분석

■ 모형 진단(다중공선성)

```
> vif(model2)
```

|              | cyl       | disp      | hp       | drat     | wt        |
|--------------|-----------|-----------|----------|----------|-----------|
| library(car) | 15.373833 | 21.620241 | 9.832037 | 3.374620 | 15.164887 |
| vif(model2)  | qsec      | vs        | am       | gear     | carb      |
|              | 7.527958  | 4.965873  | 4.648487 | 5.357452 | 7.908747  |

- cyl 변수 약 15, disp 변수 약 21, wt 변수 약 15로 높게 나타남

\*\*  $VIF_j = \frac{1}{1-R_j^2}$ ,  $R_j^2$  : 특정 독립변수  $X_j$ 를 나머지 독립변수들로 회귀분석 했을 때의 결정 계수

- 다중 공선성이 있는 것으로 보임 ( 대략 15~20정도면 다중공선성이 있는 것으로 간주함)
- 다중공선성은 독립변수들끼리 강한 상관관계를 가지는 것으로, 회귀계수의 신뢰성이 낮아지고 모형의 해석이 어려워지는 문제가 있음(변수의 제거 등 차원 축소가 필요)

## ■ 다중 회귀분석

- 변수선택법 (전진선택, 후진제거, 단계선택)
  - 1. 전진선택법 (Forward Selection)
    - 상수항만 있는 모형에서 변수를 하나씩 추가하는 방식
  - 2. 후진제거법 (Backward Elimination)
    - 전체 모형에서 변수를 하나씩 제거해나가는 방식
  - 3. 단계선택법 (Stepwise Selection)
    - 매 시도마다 전진선택법과 후진제거법을 둘 다 고려하는 방식



### ■ 다중 회귀분석

- 변수선택법 (전진선택, 후진제거, 단계선택)

\*\* AIC (Akaike Informaiton Criterion)

- R에서는 변수선택의 기준으로 AIC를 사용
- AIC는 절대적인 기준이 있는 것은 아님

상관계수처럼 다른 모델과의 상대적 비교에 사용되며, 작을 수록 좋음

- 모델이 데이터를 잘 설명하면  $\log L$  부분이 커져서 AIC가 작아짐
- 변수가 많아질 수록 패널티( $2k$ )가 증가하여 불필요한 변수 추가를 억제함

$$AIC = -2 \log L + 2k$$

#모델의 적합도를 평가

#패널티항  
(변수가 많으면 증가)

$L$  = 모델의 우도  $k$  = 모델의 자유도

### ■ 다중 회귀분석

- 변수선택법 (전진선택, 후진제거, 단계선택)

- 1. 전진선택법 (Forward Selection)

```
null<-lm(mpg ~1 ,data=mtcars) #상수항만 있는 모델
```

```
full<-lm(mpg ~ .,data=mtcars) #model2와 동일 #모든 변수가 있는 모델
```

```
summary(step(null, 1. 상수항만 있는 모델에서 출발해서
```

```
direction="forward", 2. 전진선택법으로 변수를 추가해가며
```

```
scope=list(lower=null, upper=full)) 3. 상수항만 있는 모델부터 모든 변수가 있는 모델까지 고려
)
```

## ■ 다중 회귀분석

### ■ 1. 전진선택법 (Forward Selection)

#### <Step 1>

Start: AIC=115.94  
mpg ~ 1

|        | Df | Sum of Sq | RSS     | AIC     |
|--------|----|-----------|---------|---------|
| + wt   | 1  | 847.73    | 278.32  | 73.217  |
| + cyl  | 1  | 817.71    | 308.33  | 76.494  |
| + disp | 1  | 808.89    | 317.16  | 77.397  |
| + hp   | 1  | 678.37    | 447.67  | 88.427  |
| + drat | 1  | 522.48    | 603.57  | 97.988  |
| + vs   | 1  | 496.53    | 629.52  | 99.335  |
| + am   | 1  | 405.15    | 720.90  | 103.672 |
| + carb | 1  | 341.78    | 784.27  | 106.369 |
| + gear | 1  | 259.75    | 866.30  | 109.552 |
| + qsec | 1  | 197.39    | 928.66  | 111.776 |
| <none> |    |           | 1126.05 | 115.943 |

#### <Step 2>

Step: AIC=73.22  
mpg ~ wt

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| + cyl  | 1  | 87.150    | 191.17 | 63.198 |
| + hp   | 1  | 83.274    | 195.05 | 63.840 |
| + qsec | 1  | 82.858    | 195.46 | 63.908 |
| + vs   | 1  | 54.228    | 224.09 | 68.283 |
| + carb | 1  | 44.602    | 233.72 | 69.628 |
| + disp | 1  | 31.639    | 246.68 | 71.356 |
| <none> |    |           | 278.32 | 73.217 |
| + drat | 1  | 9.081     | 269.24 | 74.156 |
| + gear | 1  | 1.137     | 277.19 | 75.086 |
| + am   | 1  | 0.002     | 278.32 | 75.217 |

#### <Step 3>

Step: AIC=63.2  
mpg ~ wt + cyl

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| + hp   | 1  | 14.5514   | 176.62 | 62.665 |
| + carb | 1  | 13.7724   | 177.40 | 62.805 |
| <none> |    |           | 191.17 | 63.198 |
| + qsec | 1  | 10.5674   | 180.60 | 63.378 |
| + gear | 1  | 3.0281    | 188.14 | 64.687 |
| + disp | 1  | 2.6796    | 188.49 | 64.746 |
| + vs   | 1  | 0.7059    | 190.47 | 65.080 |
| + am   | 1  | 0.1249    | 191.05 | 65.177 |
| + drat | 1  | 0.0010    | 191.17 | 65.198 |

### ■ 다중 회귀분석

- 1. 전진선택법 (Forward Selection)

- 최종 모형

wt, cyl, hp 3개의 변수가 선택됨

- $R^2 = 0.8431$ ,  $R^2_{adj} = 0.8263$

$R^2$  값은 감소했지만  $R^2_{adj}$  값은 증가

```
Call:
lm(formula = mpg ~ wt + cyl + hp, data = mtcars)

Residuals:
 Min 1Q Median 3Q Max
-3.9290 -1.5598 -0.5311 1.1850 5.8986

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.75179 1.78686 21.687 < 2e-16 ***
wt -3.16697 0.74058 -4.276 0.000199 ***
cyl -0.94162 0.55092 -1.709 0.098480 .
hp -0.01804 0.01188 -1.519 0.140015

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared: 0.8431, Adjusted R-squared: 0.8263
F-statistic: 50.17 on 3 and 28 DF, p-value: 2.184e-11
```

## ■ 다중 회귀분석

### ▪ 2. 후진제거법 (Backward Elimination) 최종 모형

```
summary(stepAIC(full,
 direction="backward")
)
```

#### <Step 1>

Start: AIC=70.9  
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| - cyl  | 1  | 0.0799    | 147.57 | 68.915 |
| - vs   | 1  | 0.1601    | 147.66 | 68.932 |
| - carb | 1  | 0.4067    | 147.90 | 68.986 |
| - gear | 1  | 1.3531    | 148.85 | 69.190 |
| - drat | 1  | 1.6270    | 149.12 | 69.249 |
| - disp | 1  | 3.9167    | 151.41 | 69.736 |
| - hp   | 1  | 6.8399    | 154.33 | 70.348 |
| - qsec | 1  | 8.8641    | 156.36 | 70.765 |
| <none> |    |           | 147.49 | 70.898 |
| - am   | 1  | 10.5467   | 158.04 | 71.108 |
| - wt   | 1  | 27.0144   | 174.51 | 74.280 |

#### <Step 2>

Step: AIC=68.92  
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| - vs   | 1  | 0.2685    | 147.84 | 66.973 |
| - carb | 1  | 0.5201    | 148.09 | 67.028 |
| - gear | 1  | 1.8211    | 149.40 | 67.308 |
| - drat | 1  | 1.9826    | 149.56 | 67.342 |
| - disp | 1  | 3.9009    | 151.47 | 67.750 |
| - hp   | 1  | 7.3632    | 154.94 | 68.473 |
| <none> |    |           | 147.57 | 68.915 |
| - qsec | 1  | 10.0933   | 157.67 | 69.032 |
| - am   | 1  | 11.8359   | 159.41 | 69.384 |
| - wt   | 1  | 27.0280   | 174.60 | 72.297 |

### ■ 다중 회귀분석

- 2. 후진제거법 (Backward Elimination) 최종 모형
- wt, qsec, am 3개의 변수가 선택됨

$$R^2 = 0.8497, R^2_{adj} = 0.8336$$

후진제거법 역시

$R^2$  값은 감소했지만  $R^2_{adj}$  값은 증가

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
 Min 1Q Median 3Q Max
-3.4811 -1.5555 -0.7257 1.4110 4.6610

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.6178 6.9596 1.382 0.177915
wt -3.9165 0.7112 -5.507 6.95e-06 ***
qsec 1.2259 0.2887 4.247 0.000216 ***
am 2.9358 1.4109 2.081 0.046716 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11
```

## ■ 다중 회귀분석

### ■ 3. 단계선택법 (Stepwise Selection)

```
summary(step(full,
 direction="both")
)
```

#### <Step 1>

Start: AIC=70.9  
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| - cyl  | 1  | 0.0799    | 147.57 | 68.915 |
| - vs   | 1  | 0.1601    | 147.66 | 68.932 |
| - carb | 1  | 0.4067    | 147.90 | 68.986 |
| - gear | 1  | 1.3531    | 148.85 | 69.190 |
| - drat | 1  | 1.6270    | 149.12 | 69.249 |
| - disp | 1  | 3.9167    | 151.41 | 69.736 |
| - hp   | 1  | 6.8399    | 154.33 | 70.348 |
| - qsec | 1  | 8.8641    | 156.36 | 70.765 |
| <none> |    |           | 147.49 | 70.898 |
| - am   | 1  | 10.5467   | 158.04 | 71.108 |
| - wt   | 1  | 27.0144   | 174.51 | 74.280 |

#### <Step 2>

Step: AIC=68.92  
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

|        | Df | Sum of Sq | RSS    | AIC    |
|--------|----|-----------|--------|--------|
| - vs   | 1  | 0.2685    | 147.84 | 66.973 |
| - carb | 1  | 0.5201    | 148.09 | 67.028 |
| - gear | 1  | 1.8211    | 149.40 | 67.308 |
| - drat | 1  | 1.9826    | 149.56 | 67.342 |
| - disp | 1  | 3.9009    | 151.47 | 67.750 |
| - hp   | 1  | 7.3632    | 154.94 | 68.473 |
| <none> |    |           | 147.57 | 68.915 |
| - qsec | 1  | 10.0933   | 157.67 | 69.032 |
| - am   | 1  | 11.8359   | 159.41 | 69.384 |
| + cyl  | 1  | 0.0799    | 147.49 | 70.898 |
| - wt   | 1  | 27.0280   | 174.60 | 72.297 |

### ■ 다중 회귀분석

#### ▪ 3. 단계선택법 (Stepwise Selection)

#### ▪ 최종 모형

wt, qsec, am 3개의 변수가 선택됨

$$R^2 = 0.8497, R^2_{adj} = 0.8336$$

세 방법 중 단계선택법을 가장 흔히 사용

⇒  $R^2$ 의 손실은 줄이고 3개의 변수만을 사용하는  
간단하고 효율적인 모형으로 적합

```
Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
 Min 1Q Median 3Q Max
-3.4811 -1.5555 -0.7257 1.4110 4.6610

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.6178 6.9596 1.382 0.177915
wt -3.9165 0.7112 -5.507 6.95e-06 ***
qsec 1.2259 0.2887 4.247 0.000216 ***
am 2.9358 1.4109 2.081 0.046716 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11
```



# 감사합니다

## Q&A



**충북대학교**  
CHUNGBUK NATIONAL UNIVERSITY