

# **선형회귀분석**

**1조 정해성 이지현 박소연 박병후**

# 선형회귀분석



## 회귀분석이란?

한 변수에 대하여 영향을 끼치는 다른 변수들의 관계를 함수로 나타내어 분석하는 통계적 기법

- ┌ 단순회귀분석(Simple Regression Analysis)
- └ 다중회귀분석(Multiple Regression Analysis)

특히 회귀모형에서 함수의 형태가 선형(일차식)인 경우 **선형회귀분석**이라고 한다.

## 회귀분석의 목적

1. 종속변수와 독립변수들 사이의 함수관계가 어떠한 형태를 가지고 있는지 파악
2. 종속변수에 영향을 미치는 중요한 독립변수들을 추정·검정
3. 추정된 회귀함수를 이용하여 주어진 독립변수의 값에서 종속변수의 변화를 예측하는 것

# 단순선행회귀

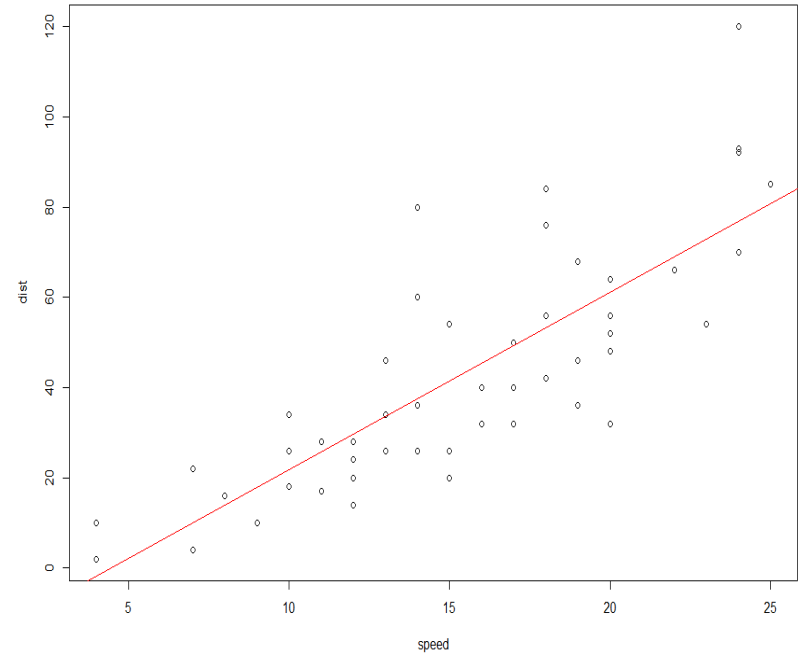
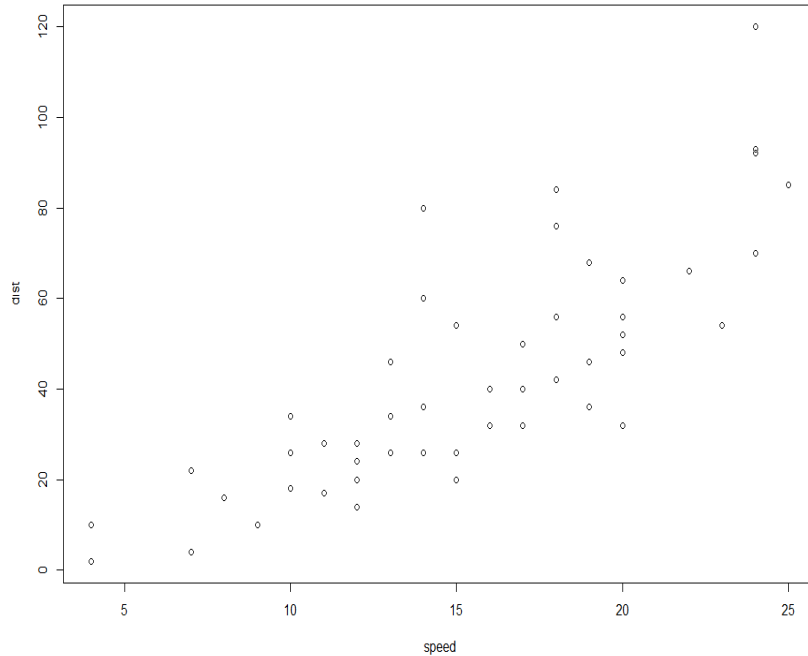
 1. 모형의 설정

 2. 모형의 적합

 3. 통계적 추론

 4. 회귀 진단 → 다음 조!

# 단순회귀 분석이란?



위와 같이 두 변수의 관계를 가장 잘 설명해주는 **선형의 관계**를 찾고 싶다.

단순회귀 분석: 하나의 독립변수  $X$ 로 양적 종속변수  $Y$ 를 예측

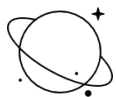
# 1. 모형의 설정



단순회귀 모형  
OLS Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$
$$i = 1, 2, \dots, n, \varepsilon_i \sim iid N(0, \sigma^2)$$

# 모형의 설정

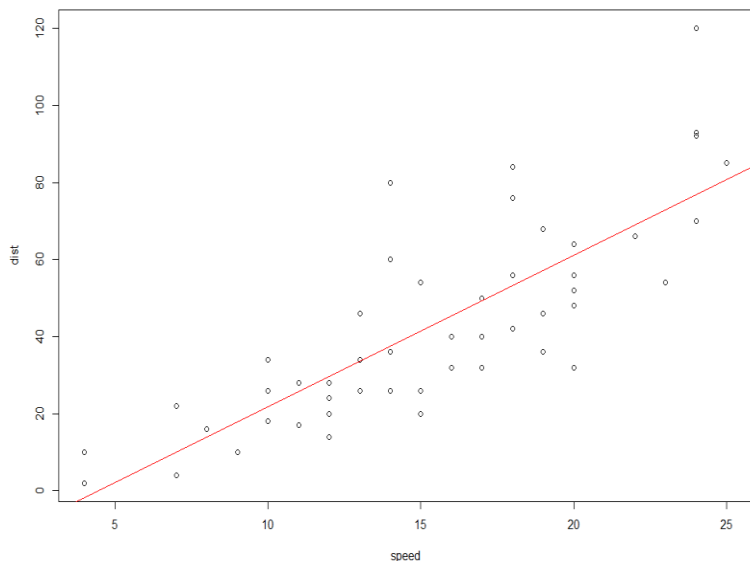


단순회귀 모형

OLS Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

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



“

$x_i$ : 설명변수, 독립변수 → 상수 취급

$y_i$ : 반응변수, 종속변수 → 확률변수  
(오차항을 포함하고 있기 때문)

$\beta_0, \beta_1$ : 모회귀계수 → 미지의 모수

$\varepsilon_i$ : 서로 독립인  $N(0, \sigma^2)$  를 따르는 오차  
→ 등분산성, 정규성, 독립성을 만족

”

실제 데이터가 생성된 시스템은 위와 같은 모형을 따른다고 가정

# 모형의 설정

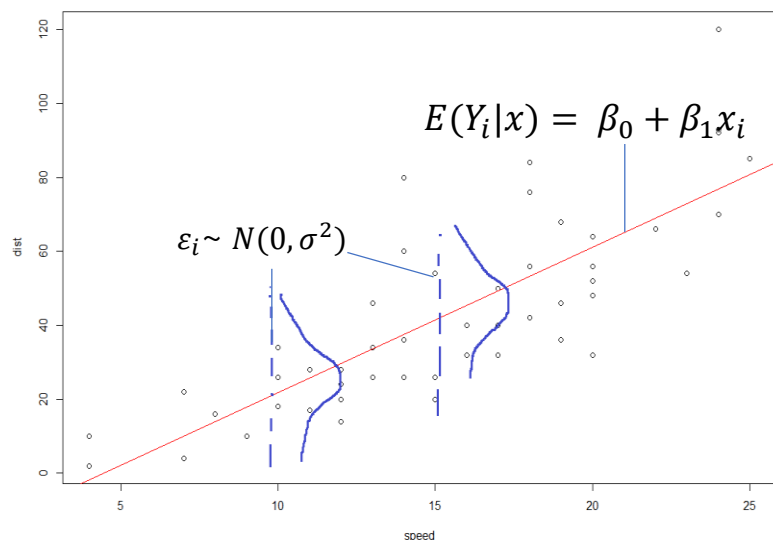


단순회귀 모형

OLS Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

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



“

**오차항의 가정**  $\varepsilon_i \sim iid N(0, \sigma^2)$

- 1)  $E(\varepsilon_i) = 0$ ,  $Var(\varepsilon_i) = \sigma^2$  (등분산성)
- 2) 오차항은 서로 독립이다. (독립성)
- 3) 오차항은 정규분포를 따른다. (정규성)

”

**회귀분석의 1차적인 목표는 표본으로부터**

**모회귀계수  $\beta_0, \beta_1$ 을 추정하여 추정된 회귀식을 만드는 것**

# 모형의 설정



단순회귀 모형

OLS Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \rightarrow \hat{y}_i = b_0 + b_1 x_i$$

회귀 모형

추정된 회귀선



## 2. 모형의 적합

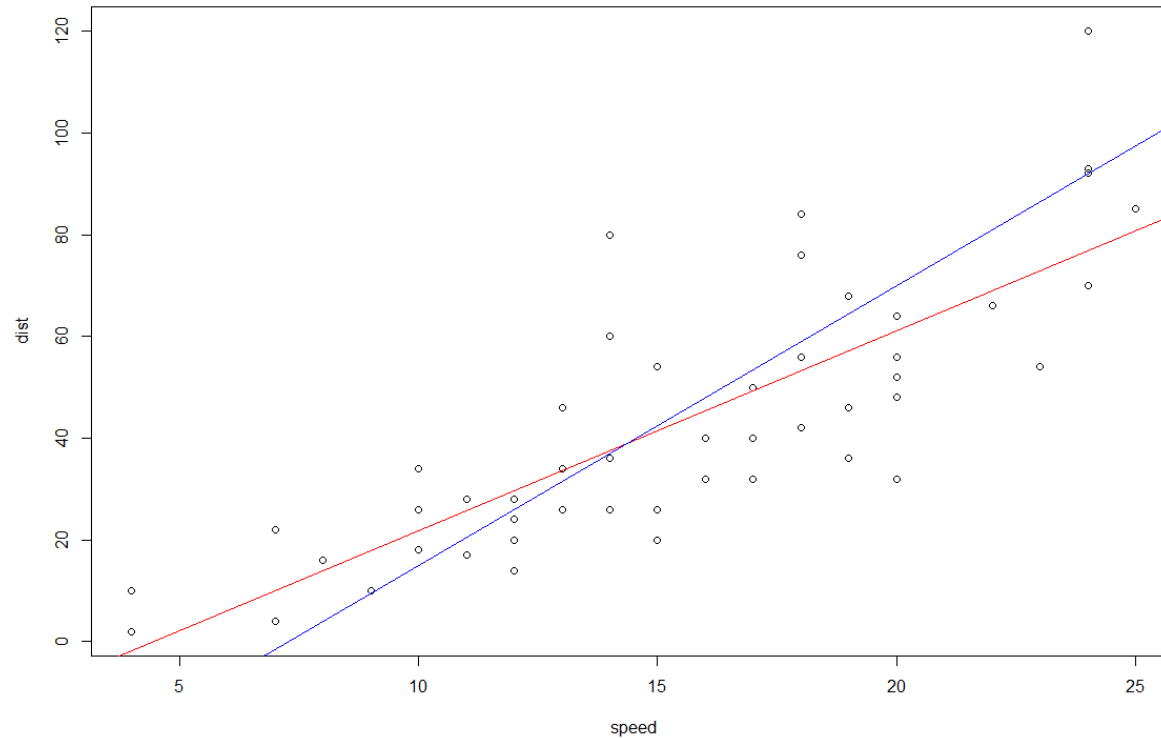


**최소제곱법**

Least Square Method

오차제곱합  $\sum_{i=1}^n \varepsilon_i^2$ 을 최소로 하는 직선을 찾는 방법

# 모형의 적합



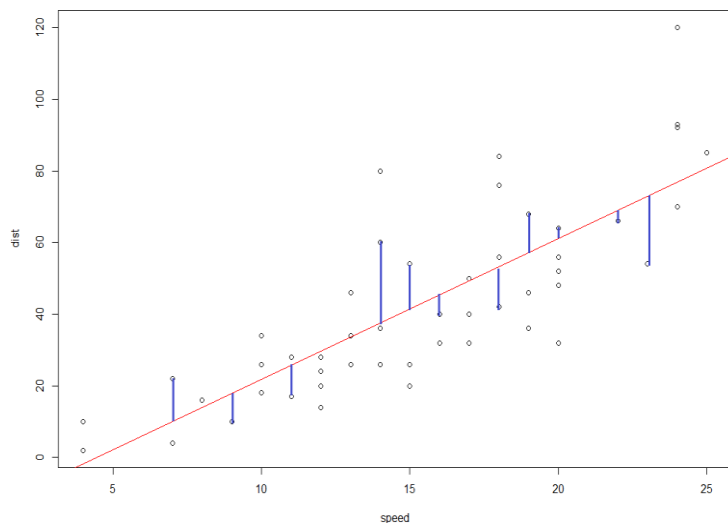
**두 개의 추정된 회귀 직선이 다음과 같을 때,  
어느 직선이 더 데이터를 잘 설명하는 직선이라고 할 수 있을까?**

# 최소제곱법



최소제곱법  
Least Square Method

오차제곱합  $\sum_{i=1}^n \varepsilon_i^2$ 을 최소로 하는 직선을 찾는 방법



“

Data  $(x_1, y_1) \cdots (x_n, y_n)$ 을 이용하여  
 $\beta_0, \beta_1$ 을 추정해보자 (각각의 추정값  $b_0, b_1$ )

최소제곱법 (Least Square Method) →  
 $\min\{ \sum_{i=1}^n \varepsilon_i^2 \} = \min\{ \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \}$

”

추정된 회귀 직선과 관측값 사이의 거리를 최소로 하는  $b_0, b_1$ 을 찾자!

# 최소제곱법



최소제곱법 (Least Square Method)  $\rightarrow \min\{\sum_{i=1}^n \varepsilon^2\} = \min\{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\}$

$$\text{Let } S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 2(-1) \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^n y_i - nb_0 - b_1 \sum_{i=1}^n x_i = 0$$

$$\therefore nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 2(-1) \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\sum_{i=1}^n x_i y_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 = 0$$

$$\therefore b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

정규방정식  
(normal equation)

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

# 최소제곱법



최소제곱법 (Least Square Method)  $\rightarrow \min\{ \sum_{i=1}^n \varepsilon^2 \} = \min\{ \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$

정규방정식  
(normal equation)

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

cf) 기호 (편의를 위해)

$$\sum (x_i - \bar{x})^2 = S(xx)$$

$$\sum (y_i - \bar{y})^2 = S(yy)$$

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

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = S(xy)$$

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

**추정된 회귀직선(by LSM):  $\hat{y}_i = b_0 + b_1 x_i$**

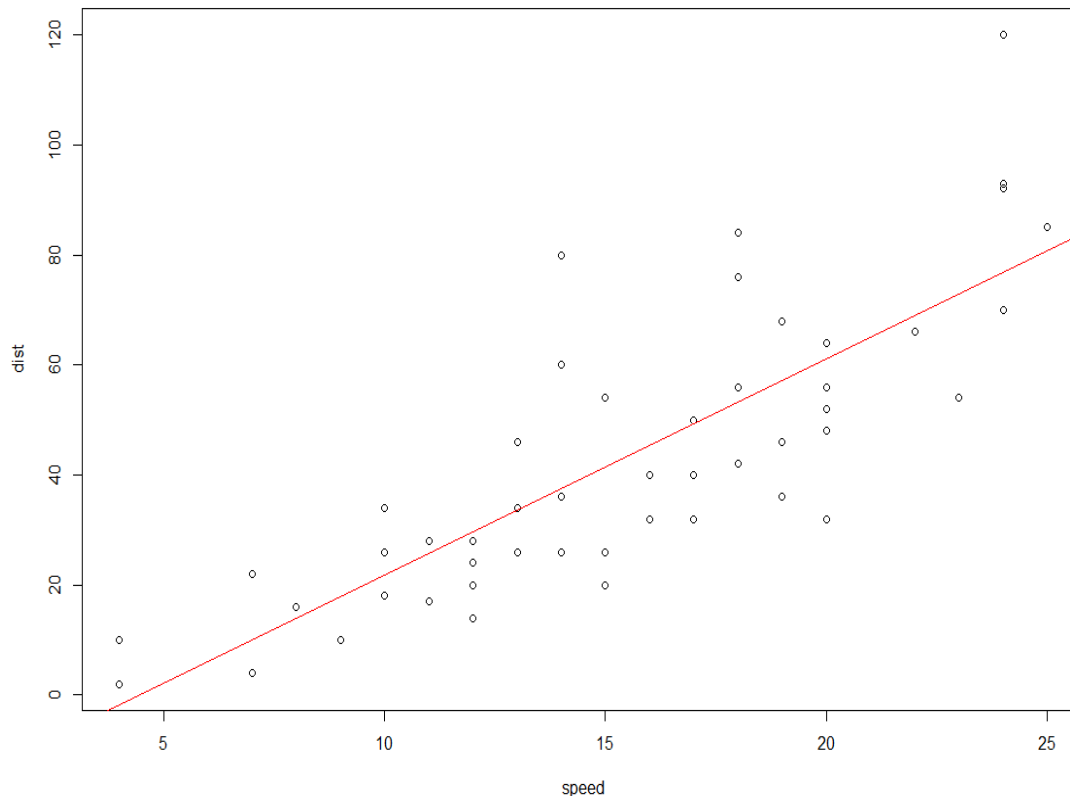
# 회귀계수의 의미 - 절편 $b_0$



회귀계수  $b_0, b_1$ 은 추정된 회귀식에서의 절편과 기울기

**절편  $b_0$ 의 의미**

→ 독립변수의 값이 0 일 때의 종속변수의 값



추정된 회귀선

$$\hat{Y} = -17.579 + 3.932X$$

```
> lm(cars$dist ~ cars$speed)

Call:
lm(formula = cars$dist ~ cars$speed)

Coefficients:
(Intercept)  cars$speed
    -17.579         3.932
```

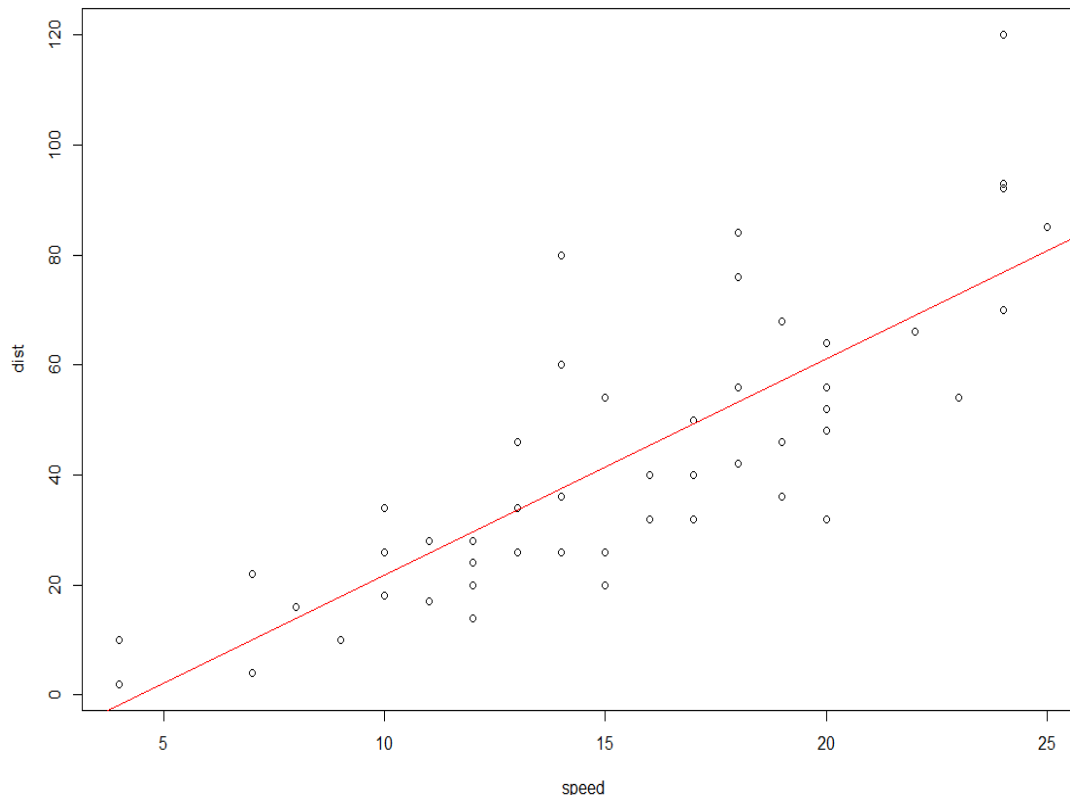
# 회귀계수의 의미 - 절편 $b_1$



회귀계수  $b_0, b_1$ 은 추정된 회귀식에서의 절편과 기울기

**기울기  $b_1$ 의 의미**

→ 독립변수  $X$ 의 값이 1단위 증가할 때 늘어나는 종속변수  $Y$ 의 값



추정된 회귀선  
 $\hat{Y} = -17.579 + 3.932X$

```
> lm(cars$dist ~ cars$speed)

Call:
lm(formula = cars$dist ~ cars$speed)

Coefficients:
(Intercept)  cars$speed
    -17.579      3.932
```

# 회귀계수의 의미 - 절편 없는 모형

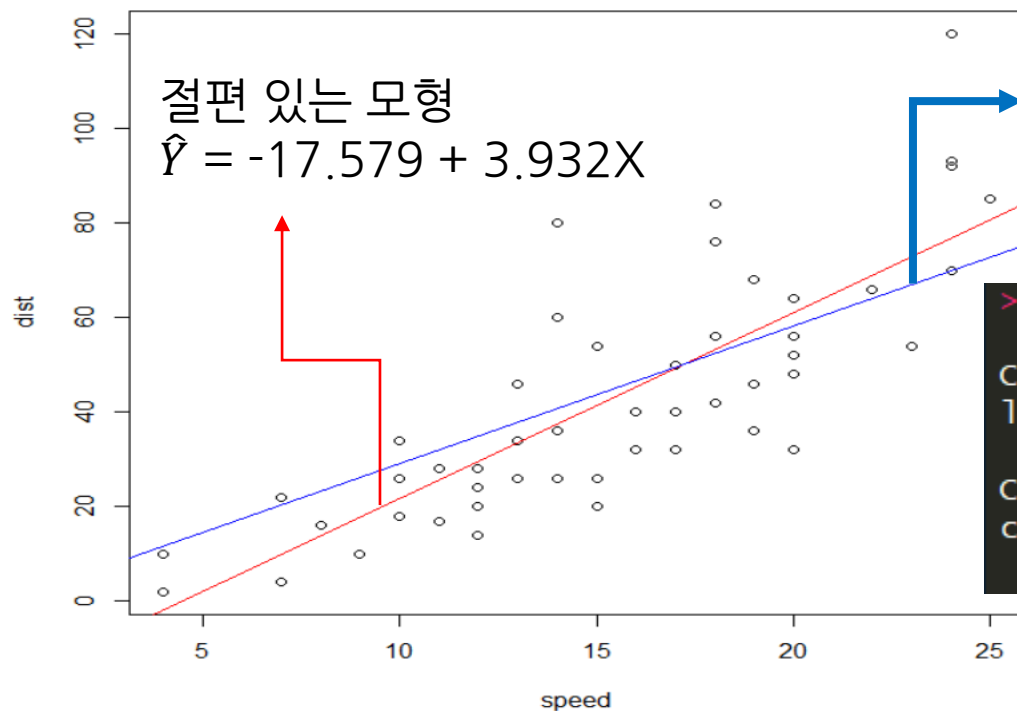


독립변수 값이 **0일 때 종속변수의 값이 반드시 0 이어야하는 자료분석**에서는

모형을 설정할 때 절편이 존재하지 않는 ‘**절편이 없는 회귀 모형**’

예) 키와 몸무게의 관계 → 키가 0이면 몸무게도 0이므로 절편이 없는 회귀모형 고려

그러나, 소득과 지출의 관계 → 소득이 없어도 지출이 있을 수 있으므로 절편이 포함된 모형



```
> lm(cars$dist ~ -1+cars$speed)

Call:
lm(formula = cars$dist ~ -1 + cars$speed)

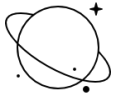
Coefficients:
cars$speed
2.909
```

$\text{lm}(y \sim -1 + x)$

절편이 없는 모형  
 $\text{lm}(\text{종속} \sim -1 + \text{독립1} + \text{독립2} + \dots)$



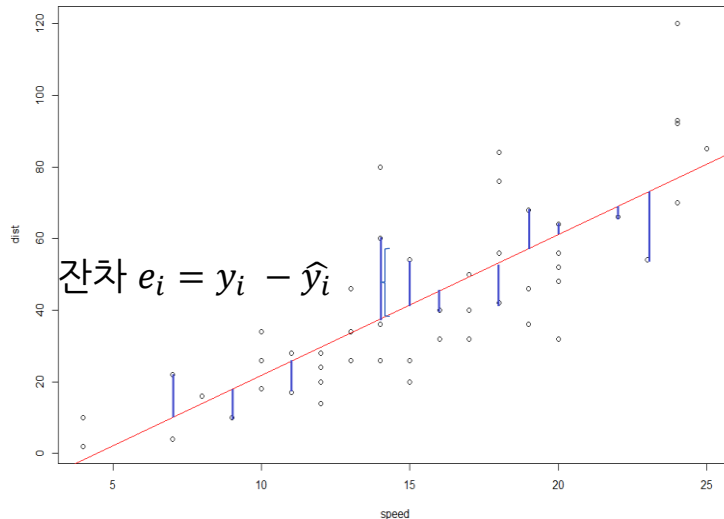
# 잔차



잔차

Residual

$$e_i = \text{실제값}(y_i) - \text{추정값}(\hat{y}_i)$$



잔차의 성질

- 1)  $\sum_{i=1}^n e_i = 0$
- 2)  $\sum_{i=1}^n x_i e_i = 0$
- 3)  $\sum_{i=1}^n \hat{y}_i e_i = 0$

cf)  $(\bar{x}, \bar{y})$ 는 추정된 회귀직선 위에 있다.

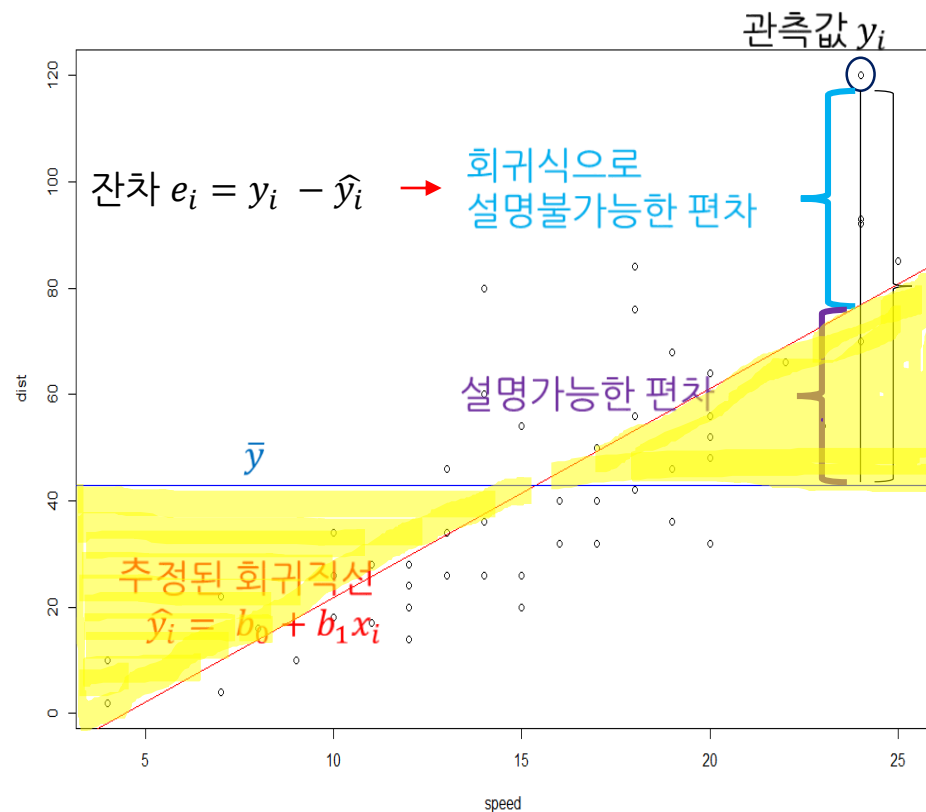
**잔차는 실제값과 추정값의 차이로 오차의 불편추정량이기 때문에 모형의 진단과정에서 중요한 역할을 한다.**

# 변동의 분해



## 분산분석표 ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



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

총편차 = 설명안되는 편차 + 설명가능편차

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

SST = SSE + SSR

pf) 양변 제곱

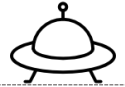
$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

여기서  $2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2 \sum_{i=1}^n e_i (y_i - \bar{y})$

$$\rightarrow 2[\sum \hat{y}_i e_i - \bar{y} \sum e_i]$$

→ 앞서 보인 잔차의 성질에 의해 0이 된다.

# 변동의 분해



## 분산분석표 ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



요인	변동	자유도	제곱평균	F비
회귀	SSR	1	$MSR = SSR/1$	$MSR/MSE$
잔차	SSE	$n-2$	$MSE = SSE/(n-2)$	
총합	SST	$n-1$		

SSE가 작을수록, 또 SSR이 클수록 회귀식의 정도가 좋다.

즉, 회귀모형이 자료들을 잘 설명해 주고 있다.

따라서 SSE , SSR/SSE등을 정도의 척도로 사용할 수 있다



# 분산분석표를 이용한 검정



분산분석표  
ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



요인	변동	자유도	제곱평균	F비
회귀	SSR	1	MSR=SSR/1	MSR/MSE
잔차	SSE	n-2	MSE=SSE/n-2	
총합	SST	n-1		

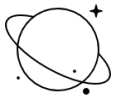
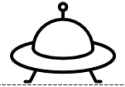
1. 결정계수 R-Squared → 모형의 설명력

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$0 \leq R^2 \leq 1$  → 1에 가까울수록 설명력이 좋은 모형  
단순선형회귀의 경우 결정계수는 상관계수의 제곱



# 분산분석표를 이용한 검정



분산분석표  
ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



요인	변동	자유도	제곱평균	F비
회귀	SSR	1	$MSR = SSR/1$	$MSR/MSE$
잔차	SSE	$n-2$	$MSE = SSE/(n-2)$	
총합	SST	$n-1$		

결정계수의 값이 1에 가까울수록 추정된 회귀식 주위에 자료가 밀집 → 자료를 잘 대표

주의) 독립변수의 수가 증가하면 결정계수는 항상 증가  
그러므로, 결정계수의 기준으로 하면 독립변수의 수가 무조건 많으면 좋다?

- 독립변수가 증가할 시 다중공선성 등의 문제 발생
- 수정결정계수 등 여러 측도를 복합적으로 사용해야 함



# 분산분석표를 이용한 검정



## 분산분석표 ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



요인	변동	자유도	제곱평균	F비
회귀	SSR	1	MSR=SSR/1	MSR/MSE
잔차	SSE	n-2	MSE=SSE/n-2	
총합	SST	n-1		

2. 유의성 검정 → 회귀선의 유의성을 검정

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_0: \beta_1 \neq 0$$

검정통계량:  $F_0 = \frac{MSR}{MSE} \sim F_{1,n-2} \quad (\text{under } H_0)$

기각역:  $F_0 > F_{1,n-2,\alpha} \quad (\text{우측검정})$



# 분산분석표를 이용한 검정



분산분석표  
ANOVA Table

자료의 변동을 회귀직선으로 설명가능한 변동과 설명 불가능한 변동으로 나누자.



요인	변동	자유도	제곱평균	F비
회귀	SSR	1	MSR=SSR/1	MSR/MSE
잔차	SSE	n-2	MSE=SSE/(n-2)	
총합	SST	n-1		

$$3. E(MSE) = \sigma^2$$

- MSE는  $\sigma^2$ 의 불편추정량(Unbiased estimator)
- 후에 회귀계수의 검정 등에서 분산의 추정량으로 쓰인다.



### 3. 통계적 추론





# 통계적 추론



단순회귀 모형  
OLS Model

기억하기

$x_i$ : 설명변수, 독립변수 → 상수

(측정오차가 없다)

$y_i$ : 반응변수, 종속변수 → 확률변수

(오차항을 포함하고 있기 때문)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad \varepsilon_i \sim iid N(0, \sigma^2)$$

$$E(y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i$$

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) \\ &= \text{Var}(\varepsilon_i) \quad (\text{분산의 성질 - 상수}) \\ &= \sigma^2 \end{aligned}$$

+ 정규분포의 성질

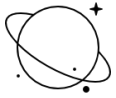
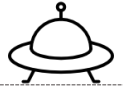
→ 정규분포에 상수를 더해도 정규분포

→  $y_i$  는 정규분포

$$\therefore y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

이를 바탕으로 통계적 추론 진행

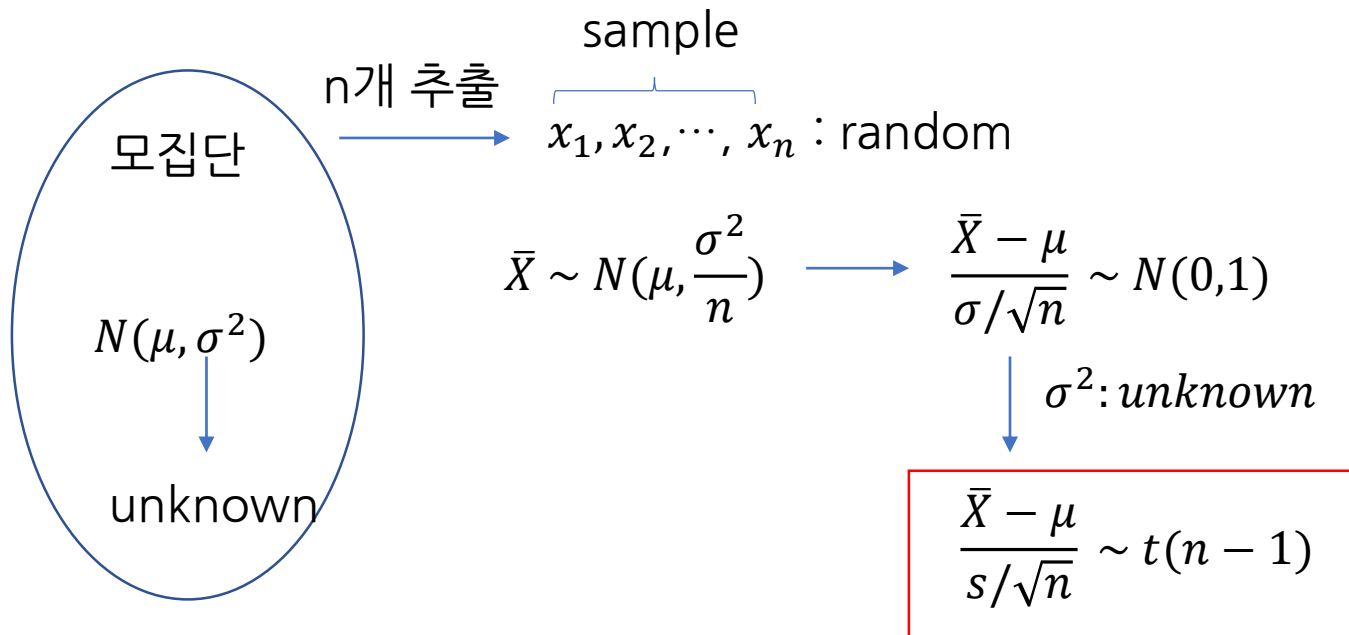
# 복습



복습

Remind

모분산  $\sigma^2$ 을 모를 때의 구간 추정



# 복습

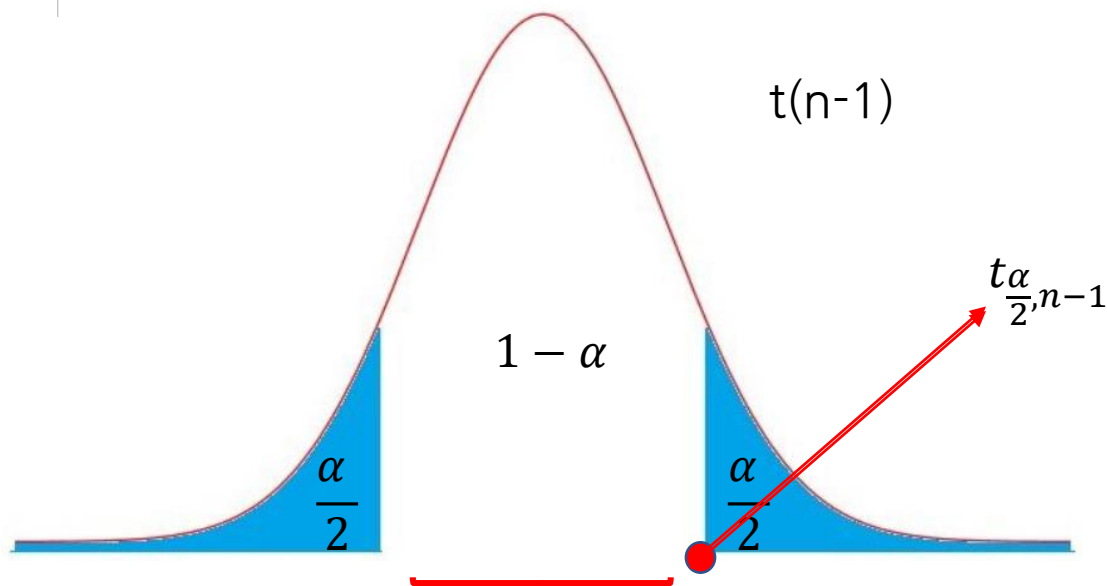


복습

Remind

모분산  $\sigma^2$ 을 모를 때의 구간 추정

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

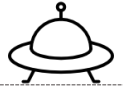


$(1 - \alpha)\%$  신뢰구간 C.I.(confidence interval)

$$\therefore P \left[ -t_{\frac{\alpha}{2}, n-1} \leq \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}, n-1} \right] = 1 - \alpha$$

$$\rightarrow \left( \bar{X} \pm t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

# $\beta_1$ 에 대한 신뢰구간



$\beta_1$ 에 대한 신뢰구간

$\beta_1$ 에 대한 불편추정량  $b_1$ 를 사용하므로  $b_1$ 의 분포를 알아야한다.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x})}{S(xx)} = \frac{\sum (x_i - \bar{x})y_i}{S(xx)} = \sum a_i y_i \quad (\text{Let } \frac{x_i - \bar{x}}{S(xx)} = a_i)$$

(편차 합 = 0)

cf)  $a_i$ 의 성질

$$\sum a_i = \sum \frac{(x_i - \bar{x})}{S(xx)} = 0 \quad (\text{편차 합} = 0)$$

$$\sum a_i^2 = \sum \frac{(x_i - \bar{x})^2}{S(xx)^2} = \frac{1}{S(xx)}$$

$$\sum a_i x_i = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{S(xx)} = \frac{S(xx)}{S(xx)} = 1$$

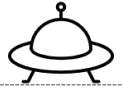
$a_i$ 는  $x$ 들의 결합

$\therefore b_1 = \sum a_i y_i \sim$  정규분포  
(정규분포의 선형결합은 정규분포 by 수리통계)

**1)  $b_1$ 은 정규분포를 따른다.**

$$\begin{aligned} (\because \sum (x_i - \bar{x})(x_i - \bar{x}) &= \sum (x_i - \bar{x})(x_i) - \bar{x} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})(x_i)) \end{aligned}$$

# $\beta_1$ 에 대한 신뢰구간



$\beta_1$ 에 대한 신뢰구간

$\beta_1$ 에 대한 불편추정량  $b_1$ 를 사용하므로  $b_1$ 의 분포를 알아야한다.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})y_i - \bar{y} \sum (x_i - \bar{x})}{S(xx)} = \frac{\sum (x_i - \bar{x})y_i}{S(xx)} = \sum a_i y_i \quad (\text{Let } \frac{x_i - \bar{x}}{S(xx)} = a_i)$$

(편차 합 = 0)

cf)  $a_i$ 의 성질

$$\sum a_i = \sum \frac{(x_i - \bar{x})}{S(xx)} = 0 \quad (\text{편차 합} = 0)$$

$$\sum a_i^2 = \sum \frac{(x_i - \bar{x})^2}{S(xx)^2} = \frac{1}{S(xx)}$$

$$\sum a_i x_i = \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{S(xx)} = \frac{S(xx)}{S(xx)} = 1$$

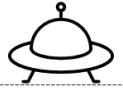
$$\begin{aligned} \mathbf{2) E(b_1)} &= E(\sum a_i y_i) = \sum a_i \cdot E(y_i) \\ &= \sum a_i (B_0 + B_1 x_i) = b_0 \sum a_i + b_1 \sum a_i x_i \\ &= \mathbf{B_1} \end{aligned}$$

$$\begin{aligned} \mathbf{3) Var(b_1)} &= Var(\sum a_i y_i) = \sum a_i^2 \cdot Var(y_i) \quad (y_i \text{ 서로 독립}) \\ &= \sum a_i^2 \cdot \sigma^2 \end{aligned}$$

$$\begin{aligned} (\because \sum (x_i - \bar{x})(x_i - \bar{x}) &= \sum (x_i - \bar{x})(x_i) - \bar{x} \sum (x_i - \bar{x}) \\ &= \sum (x_i - \bar{x})(x_i)) \end{aligned}$$

$$= \frac{\sigma^2}{S(xx)}$$

# $\beta_1$ 에 대한 신뢰구간

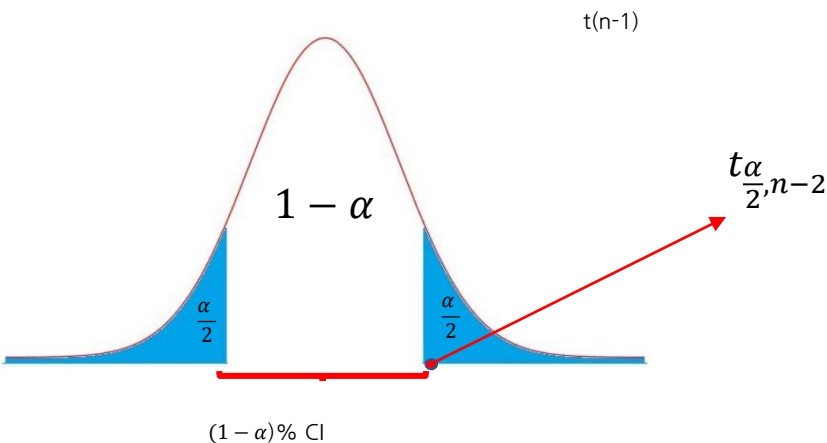


$\beta_1$ 에 대한 신뢰구간

$\beta_1$ 에 대한 불편추정량  $b_1$ 를 사용하므로  $b_1$ 의 분포를 알아야한다.

$$b_1 \sim N(B_1, \frac{\sigma^2}{S(xx)}) \quad \text{cf) } E(\text{MSE}) = \sigma^2$$

$$\therefore \frac{b_1 - B_1}{\sqrt{\frac{\sigma^2}{S(xx)}}} \sim N(0, 1) \xrightarrow[\text{unknown}]{\sigma^2} \frac{b_1 - B_1}{\sqrt{\frac{\text{MSE}}{S(xx)}}} \sim t_{(n-2)}$$



$$\therefore P \left[ -t_{\frac{\alpha}{2}, n-2} \leq \frac{b_1 - B_1}{\sqrt{\frac{\text{MSE}}{S(xx)}}} \leq t_{\frac{\alpha}{2}, n-2} \right] = 1 - \alpha$$

→  $\beta_1$ 의 (1- $\alpha$ ) % 신뢰구간

$$\rightarrow (b_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{\frac{\text{MSE}}{S(xx)}})$$

# $\beta_0$ 에 대한 신뢰구간



$\beta_0$ 에 대한 신뢰구간

$\beta_0$ 에 대한 불편추정량  $b_0$ 를 사용하므로  $b_0$ 의 분포를 알아야한다.

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

$$\bar{y} = \frac{1}{n}y_1 + \frac{1}{n}y_2 + \dots + \frac{1}{n}y_n$$

$$b_1 = a_1y_1 + a_2y_2 + \dots + a_ny_n$$

$$\sum a_i = 0$$

$y_i$ 들은 독립(오차항이 독립이므로)

**1)  $b_0$ 은 정규분포를 따른다.** (역시  $b_1$  (정규분포)의 선형결합)

$$\mathbf{2) } E(b_0) = E(\bar{y} - b_1 \bar{x}) = E(\bar{y}) - \bar{x}E(b_1)$$

$$= B_0 + B_1 \bar{x} - B_1 \bar{x} = \mathbf{B_0}$$

$$\mathbf{3) } \mathbf{Var}(b_0) = Var(\bar{y} - b_1 \bar{x}) = Var(\bar{y}) + \bar{x}^2 Var(b_1) + 2Cov(\bar{y}, b_1 \bar{x})$$

여기서  $Cov(\bar{y}, b_1 \bar{x}) = \bar{x} Cov(\bar{y}, b_1)$  이고, (공분산의 성질)

$$Cov(\bar{y}, b_1) = (\frac{1}{n}a_1 + \frac{1}{n}a_2 + \dots + \frac{1}{n}a_n) \cdot \sum Var(y_i)$$

$$= (\frac{1}{n}a_1 + \frac{1}{n}a_2 + \dots + \frac{1}{n}a_n) \cdot \sigma^2 = \frac{\sum a_i \sigma^2}{n} = 0$$

$$\therefore Var(\mathbf{b_0}) = Var(\bar{y}) - \bar{x}^2 Var(b_1)$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S(xx)}$$

참고)  $y_i$ 들이 독립일 때,

$$\text{Let, } Z_1 = c_1y_1 + c_2y_2 + \dots + c_ny_n$$

$$Z_2 = d_1y_1 + d_2y_2 + \dots + d_ny_n$$

$$Cov(Z_1, Z_2) = (c_1d_1 + c_1d_2 + \dots + c_nd_n) Var(y_1)$$

# $\beta_0$ 에 대한 신뢰구간

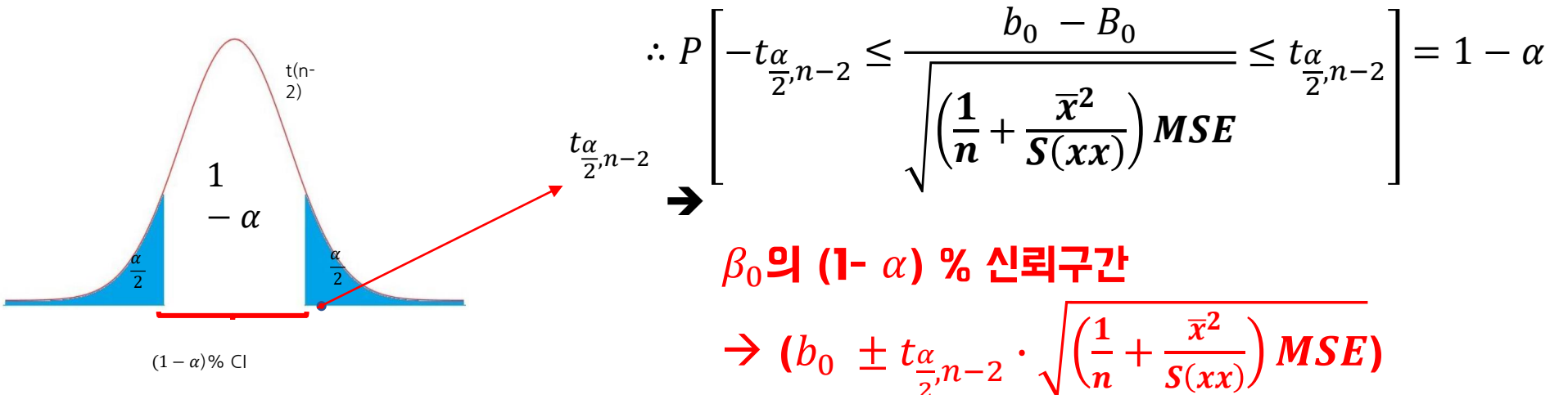


$\beta_0$ 에 대한 신뢰구간

$\beta_0$ 에 대한 불편추정량  $b_0$ 를 사용하므로  $b_0$ 의 분포를 알아야한다.

$$b_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)\sigma^2\right) \quad \text{cf) } E(MSE) = \sigma^2$$

$$\therefore \frac{b_0 - B_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)\sigma^2}} \sim N(0,1) \xrightarrow[\text{unknown}]{\sigma^2} \frac{b_0 - B_0}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)MSE}} \sim t_{(n-2)}$$





# 정리



정리

$\beta_1$ 과  $\beta_0$ 에 대한 신뢰구간과  $b_1$ ,  $b_0$ 의 분포

$$b_1 \sim N(B_1, \frac{\sigma^2}{S(xx)})$$
$$b_0 \sim N(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right) \sigma^2)$$

cf)  $E(MSE) = \sigma^2$

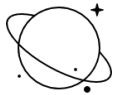
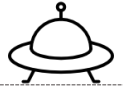
$\beta_1$ 의  $(1-\alpha)$  % 신뢰구간

$$\rightarrow (b_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{\frac{MSE}{S(xx)}})$$

$\beta_0$ 의  $(1-\alpha)$  % 신뢰구간

$$\rightarrow (b_0 \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right) MSE})$$

# $\beta_1$ 의 검정



## $\beta_1$ 의 검정

$\beta_1$ 에 대한 불편추정량  $b_1$ 를 사용하므로  $b_1$ 의 분포를 알아야한다.

$$b_1 \sim N(B_1, \frac{\sigma^2}{S(xx)})$$

$$H_0: B_1 = B_{10}(\text{상수}) \text{ vs } H_1: B_1 \neq B_{10}$$

$$B_{10}(\text{주어진 상수})$$

검정통계량

$$t_0 = \frac{b_1 - B_{10}}{\sqrt{\frac{MSE}{S(xx)}}} \quad (\sim t_{(n-2)} \text{ under } H_0)$$

기각역

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

P-value

$$P[T > |t_0|], \text{ where } T \sim t_{(n-2)}$$

# $\beta_0$ 의 검정



## $\beta_0$ 의 검정

$\beta_0$ 에 대한 불편추정량  $b_0$ 를 사용하므로  $b_0$ 의 분포를 알아야한다.

$$b_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)\sigma^2\right)$$

$$H_0: B_0 = B_{00}(\text{상수}) \text{ vs } H_1: B_0 \neq B_{00}$$

$$B_{00}(\text{주어진 상수})$$

검정통계량

$$t_0 = \frac{b_0 - B_{00}}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)MSE}} \quad (\sim t_{(n-2)} \text{ under } H_0)$$

기각역

$$|t_0| > t_{\frac{\alpha}{2}, n-2}$$

P-value

$$P[T > |t_0|], \text{ where } T \sim t_{(n-2)}$$

# 데이터 소개



## BOSTON

MASS::Boston

변수	설명
medv	중간 수준의 주택 가격 (median house value)
rm	주택당 평균 방의 수 (average number of rooms per house)
age	주택평균연령 (average age of houses)
lstat	낮은 사회경제적 지위를 가진 가구의 비율 (percent of households with low socioeconomic status)
...	이하 생략

# 데이터 소개



```
> library(MASS)
> Boston <- MASS::Boston
> head(Boston)
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

```
> str(Boston)
```

'data.frame': 506 obs. of 14 variables:

```
$ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...
$ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
$ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
$ chas : int 0 0 0 0 0 0 0 0 0 0 ...
$ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
$ rm : num 6.58 6.42 7.18 7 7.15 ...
$ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
$ dis : num 4.09 4.97 4.97 6.06 6.06 ...
$ rad : int 1 2 2 3 3 3 5 5 5 5 ...
$ tax : num 296 242 242 222 222 222 311 311 311 311 ...
$ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
$ black : num 397 397 393 395 397 ...
$ lstat : num 4.98 9.14 4.03 2.94 5.33 ...
$ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

# 모형의 설정



가설검정을 위한 모형 가정

$$\text{medv} = \beta_0 + \beta_1 \cdot \text{lstat} + \varepsilon_i$$

# 모형의 적합



```
> names(Boston)
[1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"
[10] "tax"     "ptratio" "black"   "lstat"   "medv"
```

단순선형회귀모형에 적합시켜 lm.fit에 저장 후 출력  
(코드 2줄 또는 3줄)

```
Coefficients:
(Intercept)    34.55
lstat         -0.95
```

$$\Rightarrow \widehat{\text{medv}} = 34.55 - 0.95 \text{ lstat}$$

# 통계적 추론



```
> summary(lm.fit) # p값, 표준오차, R^2, F값 등 제공
```

```
Call:
```

```
lm(formula = medv ~ lstat)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-15.168	-3.990	-1.318	2.034	24.500

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	34.55384	0.56263	61.41	<2e-16 ***
lstat	-0.95005	0.03873	-24.53	<2e-16 ***

```
---
```

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

```
Residual standard error: 6.216 on 504 degrees of freedom  
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432  
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```



# 통계적 추론



$$\text{medv} = \beta_0 + \beta_1 \cdot \text{lstat} + \varepsilon_i$$

귀무가설 : 회귀선은 유의하지 못하다. ( $\beta_1 = 0$ )

대립가설 : 회귀선은 유의하다. ( $\beta_1 \neq 0$ )

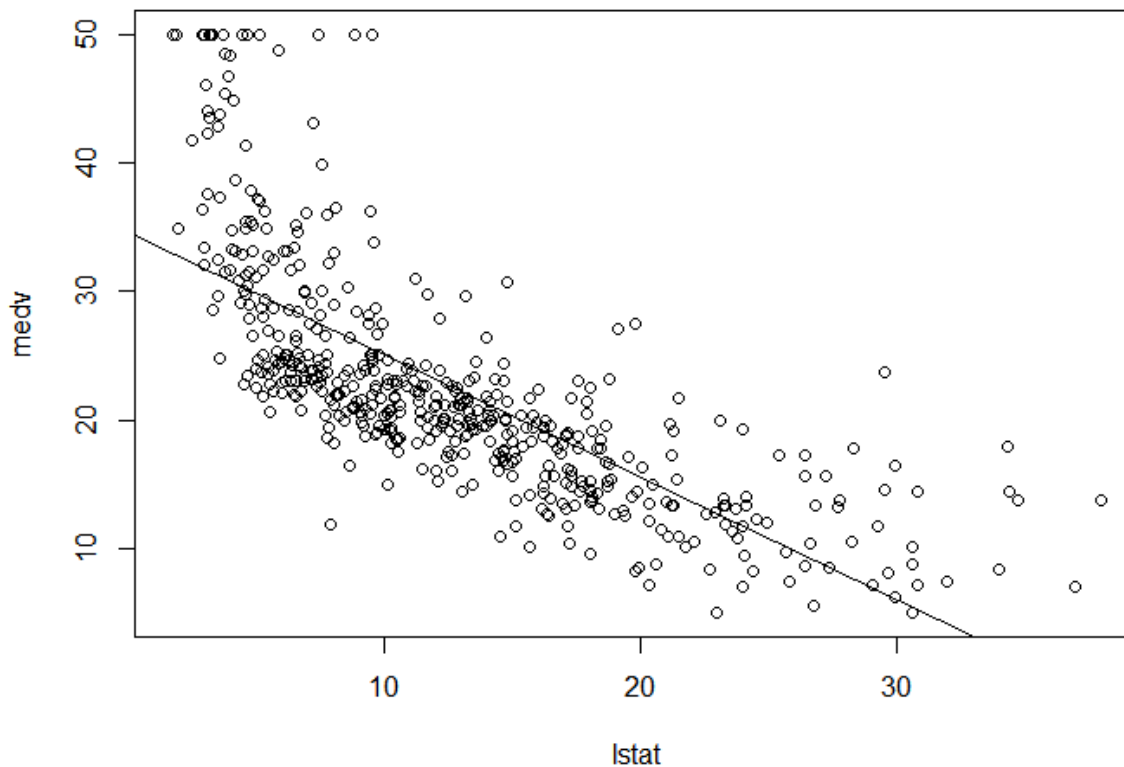
```
Residual standard error: 6.216 on 504 degrees of freedom  
Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432  
F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16
```

결정계수  $R^2$ 은 0.5441로서  
종속변수(medv) 분산의 54.4%가 독립변수(lstat)에 의해 설명됨을 알 수 있다.  
유의수준  $100 \cdot (1 - \alpha) \%$  하에서  
 $F\text{-statistic} > F_{1, n-2, \alpha}$  일 때 귀무가설을 기각할 수 있다.

# 그래프로 회귀직선 그려보기



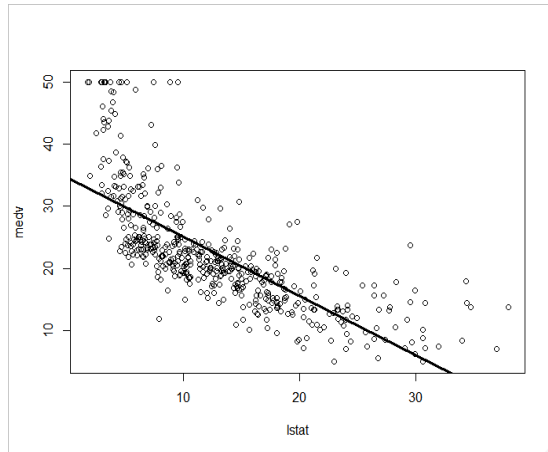
다음과 같이 그래프를 그리고 추정된 회귀직선을  
그리세요(코드 두줄)



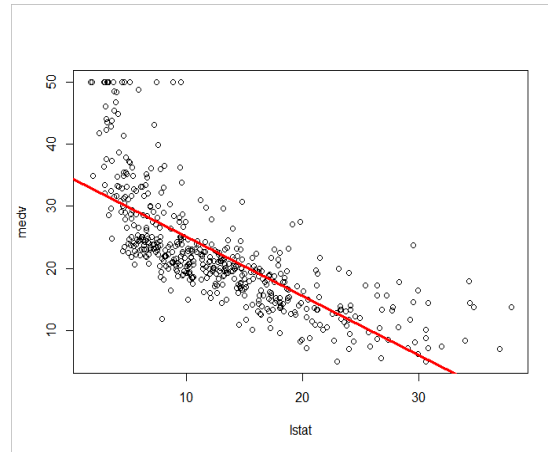
# 다양한 옵션 사용



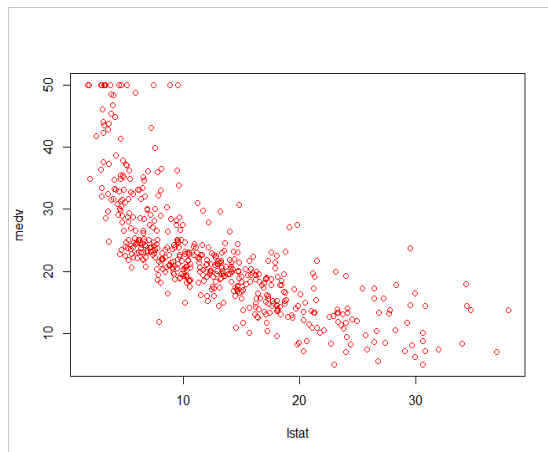
```
> abline(lm.fit, lwd=3)
```



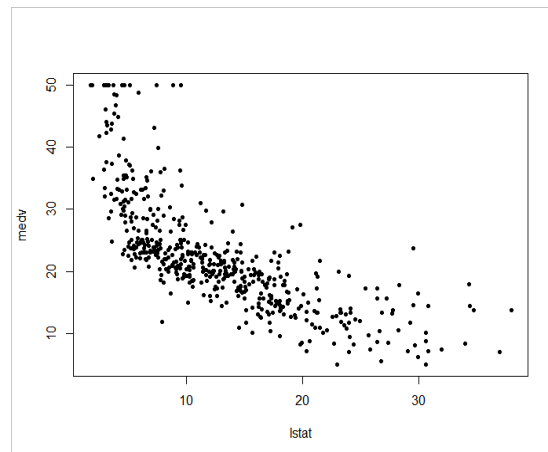
```
> abline(lm.fit, lwd=3, col="red")
```



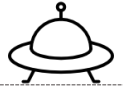
```
> plot(lstat, medv, col="red")
```



```
> plot(lstat, medv, pch=20)
```



# R의 plot 함수

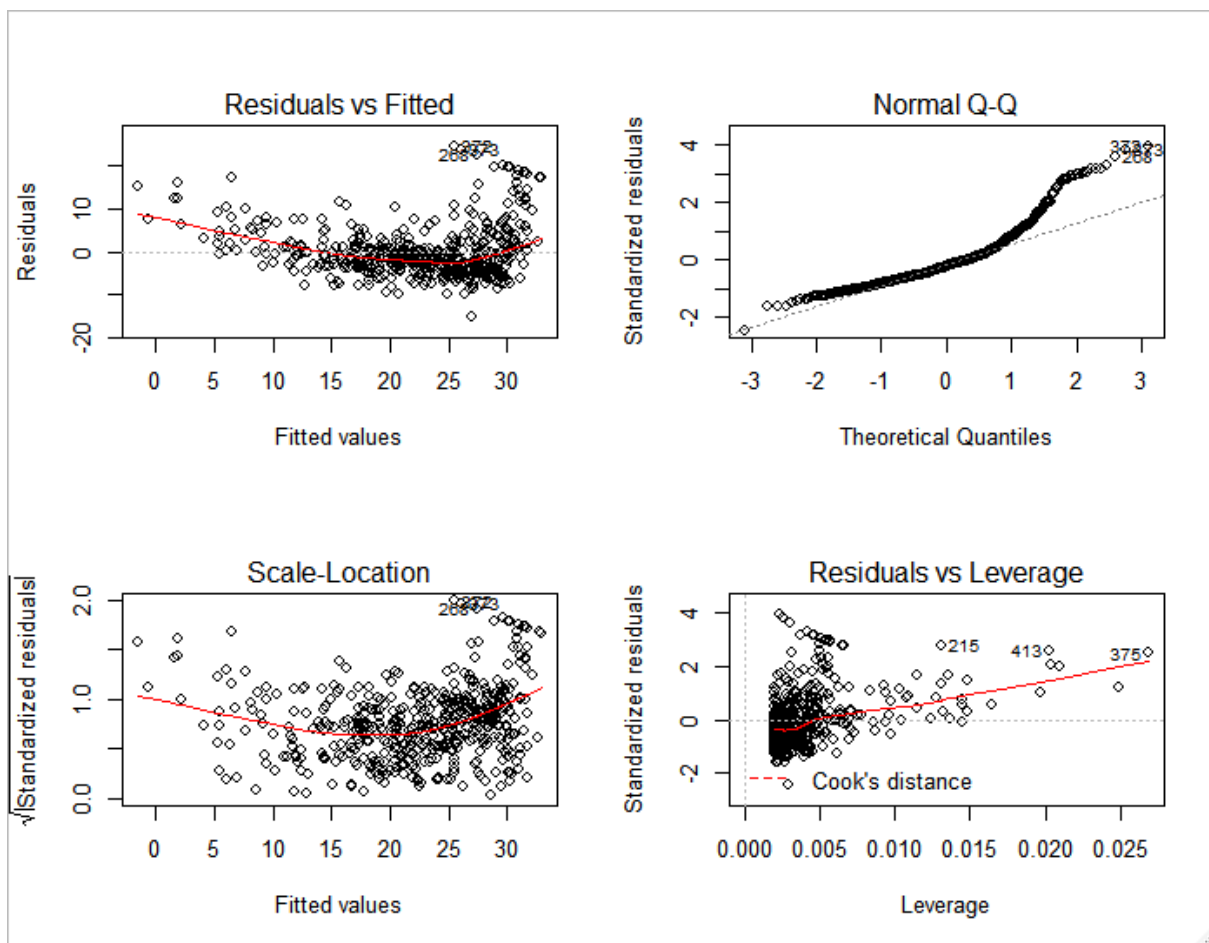


R의 plot 함수가 그린 4개의 도표는 회귀모델의 유용한 정보를 제공한다.

1. Fitted값에 대한 잔차 도표
2. 척도 위치 도표
3. 정규 Q-Q(quantile-quantile) 도표
4. 잔차와 지렛대(leverage)에 대한 도표

```
> plot(lm.fit)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> par(mfrow = c(2,2))
> plot(lm.fit)
```

# R의 plot함수





# 다중선형회귀



## 1. 가설검정



## 2. 질적예측변수



## 3. INTERACTIONS

# 다중선형회귀란?

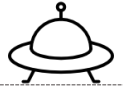


## 다중선형회귀

Multiple linear regression

다중 선형 회귀란, 설명변수  $X$ 가 여러 개인 선형회귀를 말한다.  
이때, 다변량 선형회귀(multivariate linear Regression) 과 헷갈리지  
않도록 주의한다. 다변량 선형회귀는 종속변수  $Y$ 가 여러 개인 회귀로  
여기서는 다루지 않는다.

# 다중선형회귀란?



## 다중선형회귀

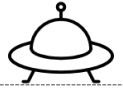
Multiple linear regression

다중 선형 회귀란, 설명변수  $X$ 가 여러 개인 선형회귀를 말한다.  
이 때, 다변량 선형회귀(multivariate linear Regression)과 헷갈리지 않도록 주의한다. 다변량 선형회귀는 종속변수  $Y$ 가 여러 개인 회귀로 여기서는 다루지 않는다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_p X_{ip} + \varepsilon_i$$



# 데이터 소개

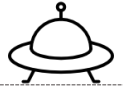


## 감독자 직무수행능력 데이터

출처 : <http://www1.aucegypt.edu/faculty/hadi/RABE5/>

변수	설명
Evaluation	상사의 직무수행에 대한 전반적인 평가
handling	피고용인의 불만 처리
privilege	특권을 허용하지 않음
opportunity	새로운 것을 배울 기회
ppro	업무 성과에 따른 승진
criticism	과실에 대한 지나친 비판
pbcc	더 나은 일로의 진급

# 모형 가정하기



가설검정을 위한 모형 가정

$$\begin{aligned} \text{Evaluation} = & \\ & \beta_0 + \beta_1 \cdot \text{handling} + \beta_2 \cdot \text{privilege} \\ & + \beta_3 \cdot \text{opportunity} + \beta_4 \cdot \text{ppro} \\ & + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon_i \end{aligned}$$

# 상관계수의 추정



단순회귀와 마찬가지로 최소제곱추정법 (LSE; 오차 제곱의 합이 최소가 되는 추정치를 찾는 방법) 을 통해 상관계수를 추정한다. 우리는 어차피 R이 해줄 것이기 때문에 생략한다.

```
> super.lm = lm(Evaluation ~ handling + privilege + opportunity + ppro + criticism + pbc , data=super)
> super.lm
```

Call:

```
lm(formula = Evaluation ~ handling + privilege + opportunity +  
    ppro + criticism + pbc, data = super)
```

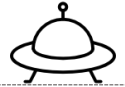
Coefficients:

(Intercept)	handling	privilege	opportunity
10.78708	0.61319	-0.07305	0.32033
ppro	criticism	pbc	
0.08173	0.03838	-0.21706	

베타값의  
추정치를  
보여주는  
부분

Beta0

# 상관계수의 추정



```
> summary(super.lm)
```

Call:

```
lm(formula = Evaluation ~ handling + privilege + opportunity +  
    ppro + criticism + pbc, data = super)
```

**Residuals:**

Min	1Q	Median	3Q	Max
-10.9418	-4.3555	0.3158	5.5425	11.5990

잔차의 분포를 보여준다. 우리는 앞선 가정( 잔차는 표준정규분포를 따른다 .)을 통해 이 모형이 적합한지 평가할 수 있다.  
여기서는 Median이 0에 가깝고, 1Q,3Q의 절댓값이 비슷하고, min, max의 절댓값 또한 비슷하므로 적절하다고 판단한다.

Coefficients:	베타추정치	s.e.(beta.hat)	가설검정부분
	Estimate	Std. Error	t value Pr(> t )
(Intercept)	10.78708	11.58926	0.931 0.361634
handling	0.61319	0.16098	3.809 0.000903 ***
privilege	-0.07305	0.13572	-0.538 0.595594
opportunity	0.32033	0.16852	1.901 0.069925 .
ppro	0.08173	0.22148	0.369 0.715480
criticism	0.03838	0.14700	0.261 0.796334
pbc	-0.21706	0.17821	-1.218 0.235577

---

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

Sqrt(MSE)  
**Residual standard error:** 7.068 on 23 degrees of freedom SSE의 자유도는  $n - p - 1$   
Multiple R-squared: 0.7326, Adjusted R-squared: 0.6628  
F-statistic: 10.5 on 6 and 23 DF, p-value: 1.24e-05

# 1. 가설검정

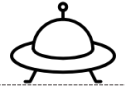


**다중선형회귀**

Multiple linear regression

가설검정시에 행렬을 통해 가설을 검정하는 경우도 있지만, 그건 시간이 부족하기때문에 생략하고, 다양한 가설에 대해서 소개합니다. 다중선형회귀의 가설검정은 주로 어떤 회귀모형을 적합할지를 결정할 때 사용합니다.

# Adjusted R-squared



다중회귀에서는, SST, SSE, SSR을 대신해서 MST, MSE, MSR을 사용한다.  
SS-가 제곱합이기 때문에, 결정계수 R square는 predictor가 많아질수록 커지기 때문이다. 따라서 SS-를 그 값의 자유도로 나눠줘서 값을 보정해준다.

Adjusted R-squared: 0.6628

요인	제곱합	자유도	평균제곱	F - statistic	Adjusted R-squared
회귀	SSR	p	MSR	$\frac{MSR}{MSE}$	$1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST}$
잔차	SSE	n-p-1	MSE		
평균	SST	n-1	MST		

모형에 predictor가 많아질수록, SSE의 값은 작아진다. 따라서 R<sup>2</sup>의 값이 1에 가까워진다. 즉 무조건 변수 X를 많이 적합하면 적절한 모형이라는 잘못된 결과가 나올 수 있으므로, 자유도 n(관측값수) - p(predictor 개수) - 1로 SSE를 나눠줘서 보정해주는 것이다!

# 전체모형의 적합도 추정



이제 앞서 적합한 모형 (Full Model) 의 적합도를 판단해 볼 것이다. 이때는 F 분포값을 이용해서 검정하는데, 아까의 summary를 통해 판단할 수 있다.

이때,

귀무가설( $H_0$ ):  $\beta_0 = \beta_1 = \dots \beta_6 = 0$

대립가설:  $\beta$  중 적어도 하나는 0이 아니다.

이고, 유의수준  $100 \times (1 - \alpha) \%$  하에서

$F\text{-statistic} > F_{(p, n-p-1, 1-\alpha)}$  일 때  
귀무가설을 기각할 수 있다.

( $p = 6, n-p-1 = 23$ )

```
> summary(super.lm)

Call:
lm(formula = Evaluation ~ handling + privilege + opportunity +
    ppro + criticism + pbc, data = super)

Residuals:
    Min       1Q   Median       3Q      Max
-10.9418  -4.3555   0.3158   5.5425  11.5990

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.78708    11.58926   0.931 0.361634
handling      0.61319     0.16098   3.809 0.000903 ***
privilege    -0.07305     0.13572  -0.538 0.595594
opportunity   0.32033     0.16852   1.901 0.069925 .
ppro          0.08173     0.22148   0.369 0.715480
criticism     0.03838     0.14700   0.261 0.796334
pbc          -0.21706     0.17821  -1.218 0.235577
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared:  0.7326,    Adjusted R-squared:  0.6628
F-statistic: 10.5 on 6 and 23 DF, p-value: 1.24e-05
```

# 1 개 혹은 그보다 더 많은 변수에 대한 가설검정



전체모형이 얼마나 적합한지 판단하는 것 외에도, 각각의 변수가 유의한지 아닌지를 검정할 수도 있다.  
이때는 Full model 과 Reduced model 의 비교를 통해서 가설을 검정한다.  
이 검정으로 적합한 모델을 고를 수 있다.

귀무가설 :  $\beta_2 = \beta_3 = 0$

대립가설 :  $\beta_2$  와  $\beta_3$  둘 중 적어도 하나는 0이 아니다.

FM : Evaluation =  $\beta_0 + \beta_1 \cdot \text{handling} + \beta_2 \cdot \text{privilege} + \beta_3 \cdot \text{opportunity} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

RM : Evaluation =  $\beta_0 + \beta_1 \cdot \text{handling} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

$$\mathbf{F \text{ STATISTIC}} = \frac{SSE(RM) - SSE(FM) / k}{SSE(FM) / n - p - 1} \sim F_{dfR - dfF, dfF}$$

$$K = DF(RM) - DF(FM)$$



# 1 개 혹은 그보다 더 많은 변수에 대한 가설검정



이때 두 모델간의 비교는 F test를 통해 진행한다. 따라서 f test에 대한 함수를 만들어본다.  
앞 슬라이드에서 설명한 F값을 계산하고 , p.value를 내뱉는 함수를 만들어보자.

```
f.test.lm = function(R.lm, F.lm) {  
  SSE.R = sum(resid(R.lm)^2)  
  SSE.F = sum(resid(F.lm)^2)  
  df.num = R.lm$df - F.lm$df  
  df.den = F.lm$df  
  F = ((SSE.R - SSE.F) / df.num) / (SSE.F / df.den)  
  p.value = 1 - pf(F, df.num, df.den)  
  return(data.frame(F, df.num, df.den, p.value))  
}
```

# 1 개 혹은 그보다 더 많은 변수에 대한 가설검정



귀무가설 :  $\beta_2 = \beta_3 = 0$

대립가설 :  $\beta_2$  와  $\beta_3$  둘 중 적어도 하나는 0이 아니다.

FM : Evaluation =  $\beta_0 + \beta_1 \cdot \text{handling} + \beta_2 \cdot \text{privilege} + \beta_3 \cdot \text{opportunity} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

RM : Evaluation =  $\beta_0 + \beta_1 \cdot \text{handling} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

```
> super1 <- super%>%  
+   select( -c(privilege,opportunity))  
> super1.lm = lm(Evaluation ~ handling + ppro + criticism + pbc ,data=super1)  
> f.test.lm(super1.lm, super.lm)  
      F df.num df.den  p.value  
1 1.846191    2    23 0.1804745
```

Reduced Model의 F 통계량은 1.846191 이고 p-value 는 0.1804745이다.

따라서  $\alpha = 0.05$  일때 p-value가  $\alpha$  보다 크므로 귀무가설을 기각하지 못한다.

즉, privilege 변수와 opportunity는 Evaluation에 유의한 영향을 주지 못하므로  
이 두개를 빼고 모형을 적합하는게 좋다.

# 어떤 변수들의 상관계수가 서로 같은지 검정



귀무가설 :  $\beta_1 = \beta_5 (= \beta^*)$

대립가설 :  $\beta_1 \neq \beta_5$

FM :  $\text{Evaluation} = \beta_0 + \beta_1 \cdot \text{handling} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

RM :  $\text{Evaluation} = \beta_0 + \beta^* \cdot \text{handling} + \beta_4 \cdot \text{ppro} + \beta^* \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

즉,

RM:  $\text{Evaluation} = \beta_0 + \beta^* \cdot (\text{handling} + \text{criticism}) + \beta_4 \cdot \text{ppro} + \beta_6 \cdot \text{pbc} + \varepsilon$

새로운 변수  $\text{handling} + \text{criticism} = \text{hdcri}$  를 형성한 후 F- test 해준다.

# 어떤 변수들의 상관계수가 서로 같은지 검정



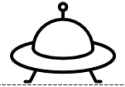
```
super2<- super1 %>%  
  mutate( hdcri = handling + criticism )  
  
superR2.lm = lm(Evaluation ~ hdcri + ppro + pbc , data= super2)  
  
f.test.lm(superR2.lm , super1.lm)  
  
> f.test.lm(superR2.lm , super1.lm)  
              F df.num df.den      p.value  
1 12.54351      1      25 0.001590039  
.
```

Reduced Model의 F 통계량은 12.54351 이고 p-value 는 0.001590039이다.

따라서  $\alpha = 0.05$  일때 p-value가  $\alpha$  보다 작으로 귀무가설을 기각한다.

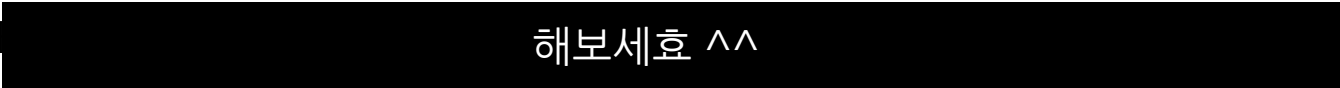
즉, handling 과 criticism은 Evaluation에 같은 수준의 상관계수를 가지고 있지 않다.

# 어떤 변수들의 상관계수가 서로 같은지 검정



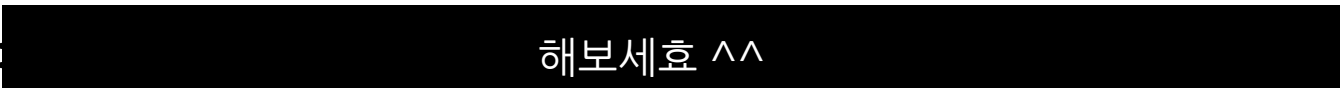
귀무가설 :  $\beta_1 = 2.5 \cdot \beta_5$   
대립가설 :  $\beta_1 \neq 2.5 \cdot \beta_5$

FM :  $\text{Evaluation} = \beta_0 + \beta_1 \cdot \text{handling} + \beta_4 \cdot \text{ppro} + \beta_5 \cdot \text{criticism} + \beta_6 \cdot \text{pbc} + \varepsilon$

RM :  해보세요 ^^

$\varepsilon$

즉,

RM :  해보세요 ^^

# 어떤 변수들의 상관계수가 서로 같은지 검정



귀무가설 :  $\beta_1 = 2.5 \cdot \beta_5$   
대립가설 :  $\beta_1 \neq 2.5 \cdot \beta_5$

해보세호 ^^

## 2. 질적예측변수



**질적예측변수**

QUALITATIVE VARIABLES

질적 예측 변수란, 성별, 거주지 처럼 카테고리화 할 수 있는 범주형 변수를 말한다. 이 변수들은 R에서는 factor 로 나타난다.  
이 챕터에서는 이 factor 변수들을 어떻게 회귀모형에 적합할지 배워보자!

# 데이터 소개



## SALARY

출처 : <http://www1.aucegypt.edu/faculty/hadi/RABE5/>

변수	설명
S (=Y)	IT 직군 고용자의 월급
X	경력 (년)
E	교육수준 (1 =학사, 2=석사 , 3= 박사)
M	관리직 유무 1= 관리직임 , 0= 관리직아님



# 데이터 불러오기



우선, 데이터를 불러와서 검사해준다.

```
load("salary.Rdata")
str(salary.table)|
salary.table$E <- factor(salary.table$E)
salary.table$M <- factor(salary.table$M)
head(salary.table$E)
```

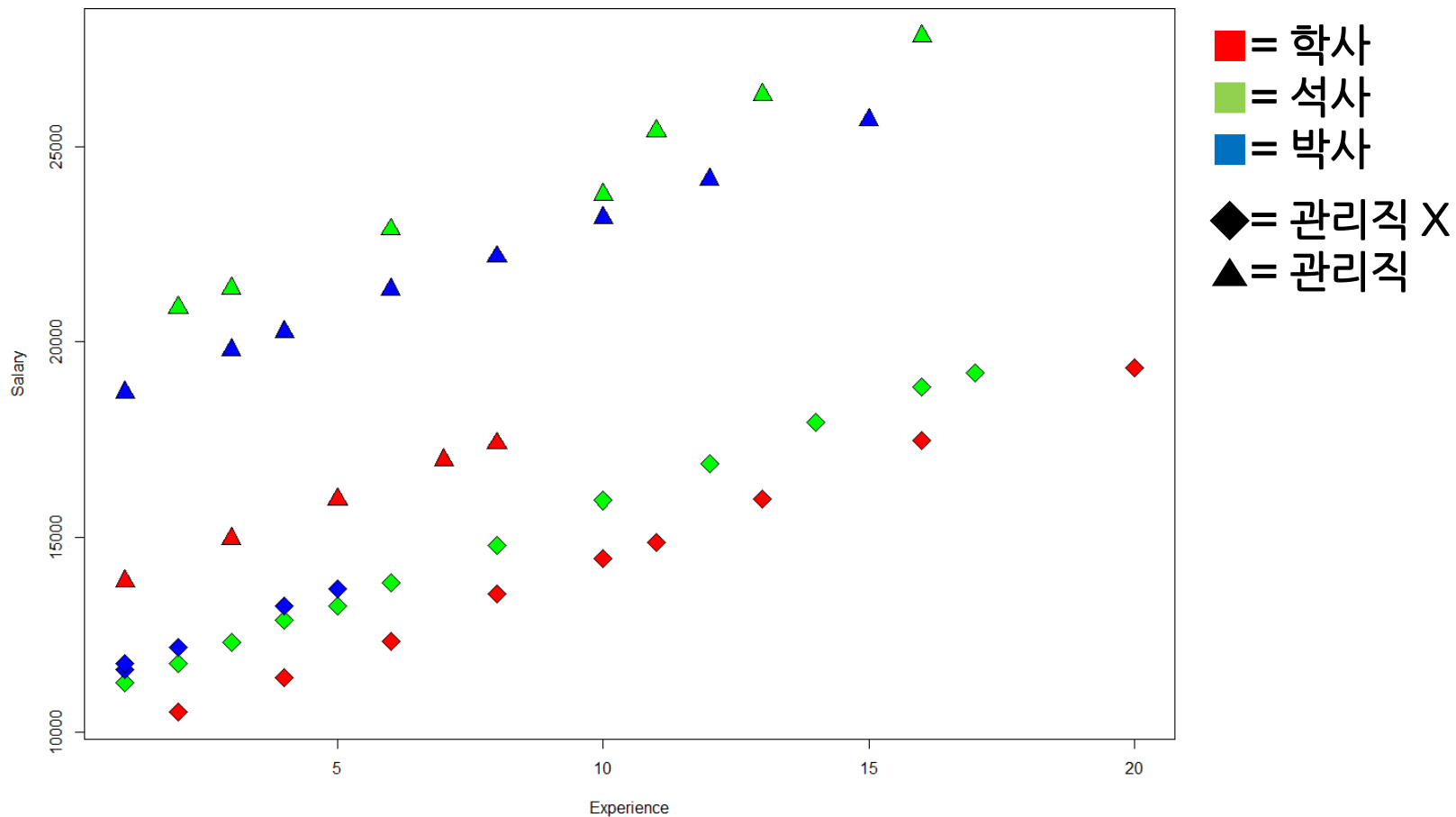
```
> str(salary.table)
'data.frame': 46 obs. of 4 variables:
 $ S: int 13876 11608 18701 11283 11767 20872 11772 10535 12195 12313 ...
 $ X: int 1 1 1 1 1 2 2 2 2 3 ...
 $ E: Factor w/ 3 levels "1","2","3": 1 3 3 2 3 2 2 1 3 2 ...
 $ M: Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 1 1 1 ...
> head(salary.table$E)
[1] 1 3 3 2 3 2
Levels: 1 2 3
```

# 시각적 판단

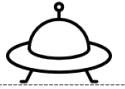


```
plot(salary.table$X, salary.table$S, type='n', xlab='Experience', ylab='Salary')
colors <- c('red', 'green', 'blue')
symbols <- c(23,24)
for (i in 1:3) {
  for (j in 0:1) {
    subset <- as.logical((salary.table$E == i) * (salary.table$M == j))
    points(salary.table$X[subset], salary.table$S[subset], pch=symbols[j+1], bg=colors[i], cex=2)
  }
}
```

# 시각적 판단



# 모형 가정하기

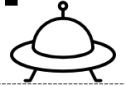


가설검정을 위한 모형 가정

$$S = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot E + \beta_3 \cdot M + \varepsilon_i$$

질적예측변수를 어떻게 모형에  
적용할까?

# 질적예측변수를 어떻게 모형에 적용할까?



정말로 친절하게도, R은 자동으로 해줍니다 ^\_^ ~!

```
salary.lm <- lm(S ~ E + M + X, salary.table)
summary(salary.lm)
```

```
> summary(salary.lm)

Call:
lm(formula = S ~ E + M + X, data = salary.table)

Residuals:
    Min       1Q   Median       3Q      Max
-1884.60  -653.60   22.23   844.85  1716.47

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8035.60     386.69  20.781 < 2e-16 ***
E2             3144.04     361.97   8.686 7.73e-11 ***
E3             2996.21     411.75   7.277 6.72e-09 ***
M1             6883.53     313.92  21.928 < 2e-16 ***
X              546.18      30.52  17.896 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1027 on 41 degrees of freedom
Multiple R-squared:  0.9568,    Adjusted R-squared:  0.9525
F-statistic: 226.8 on 4 and 41 DF,  p-value: < 2.2e-16
```

# 질적예측변수를 어떻게 모형에 적용할까?



그렇다면, R은 어떤 알고리즘으로 변수를 만드는걸까?

R은 변수가 Factor 형일 때,  
그 변수의 level 수보다 1개 적은  
수의 변수를 만든다.

그 이유는, E2, E3만 있으면  
E1 그룹을 구별할 수 있기 때문!

E2=E3 =0 이면 학사(E=1)  
E2=1, E3=0 이면 석사(E=2)  
E2=0, E3=1 이면 석사(E=3)

회귀분석에서 변수가 너무 많아지면  
과적합이 일어나기때문에 정확한  
검정을 위해 변수를 최대한 적게  
적합합니다. 이 때 다른 변수가  
0일때의 그룹(학사)은 레퍼런스  
그룹이라고 합니다.

```
> head(model.matrix(salary.lm))
      (Intercept) E2 E3 M1 X
1             1  0  0  1  1
2             1  0  1  0  1
3             1  0  1  1  1
4             1  1  0  0  1
5             1  0  1  0  1
6             1  1  0  1  2
```

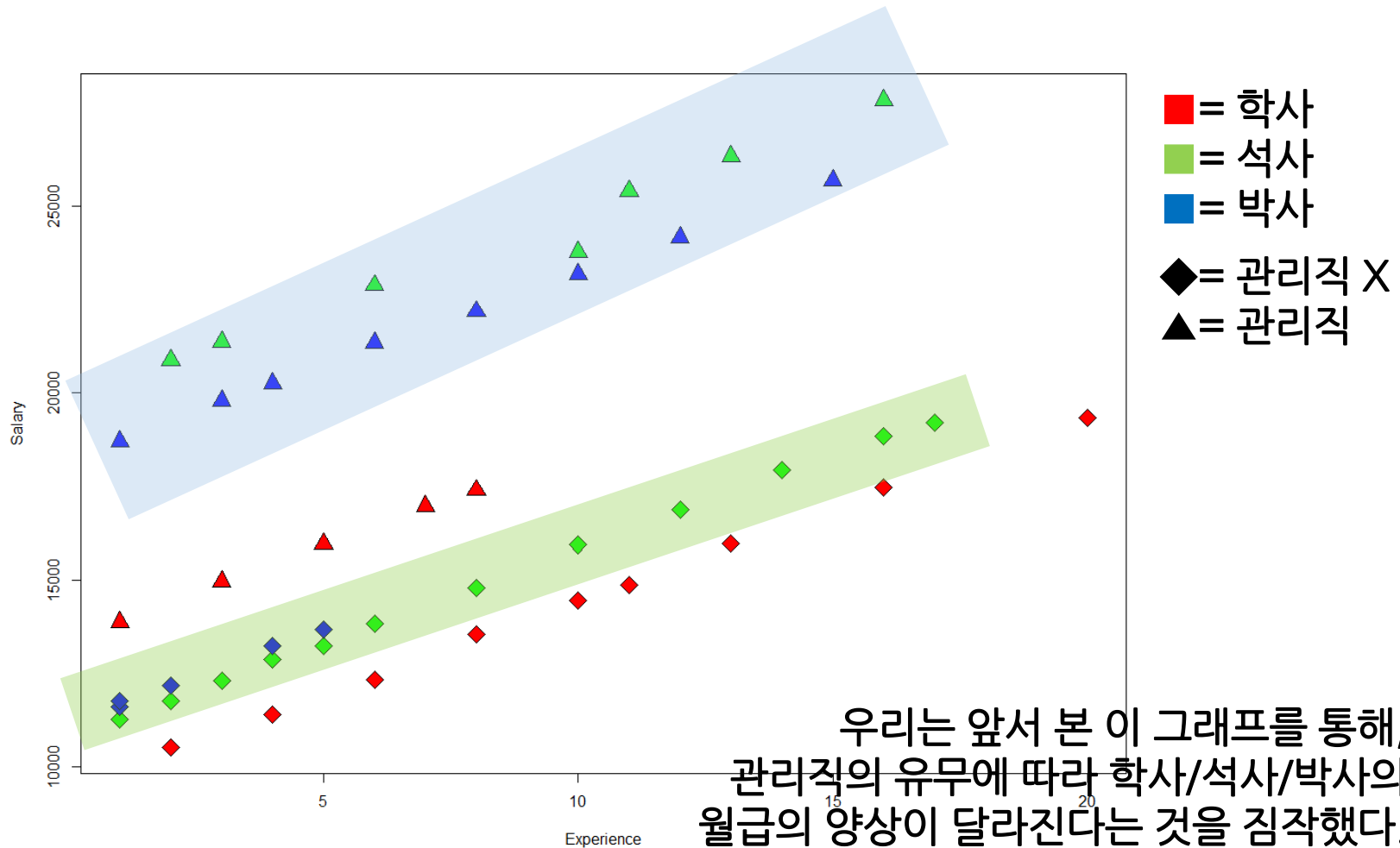
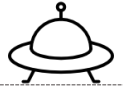
# 3. INTERACTIONS



**상호작용변수**  
INTERACTIONS

상호작용 변수란, 설명변수들 사이에 연관성이 존재할 경우를 의미한다. 하나의 변수가 다른 변수에 따라 달라지는 경우를 고려해 회귀모형을 적합할 수 있다. 따라서, <어떤 변수에 따라 어떤 변수가 달라진다>라는 가설을 검정하고 싶을 때 상호작용 변수를 사용한다. (무조건사용 X)

# 상호작용 변수가 필요할까?

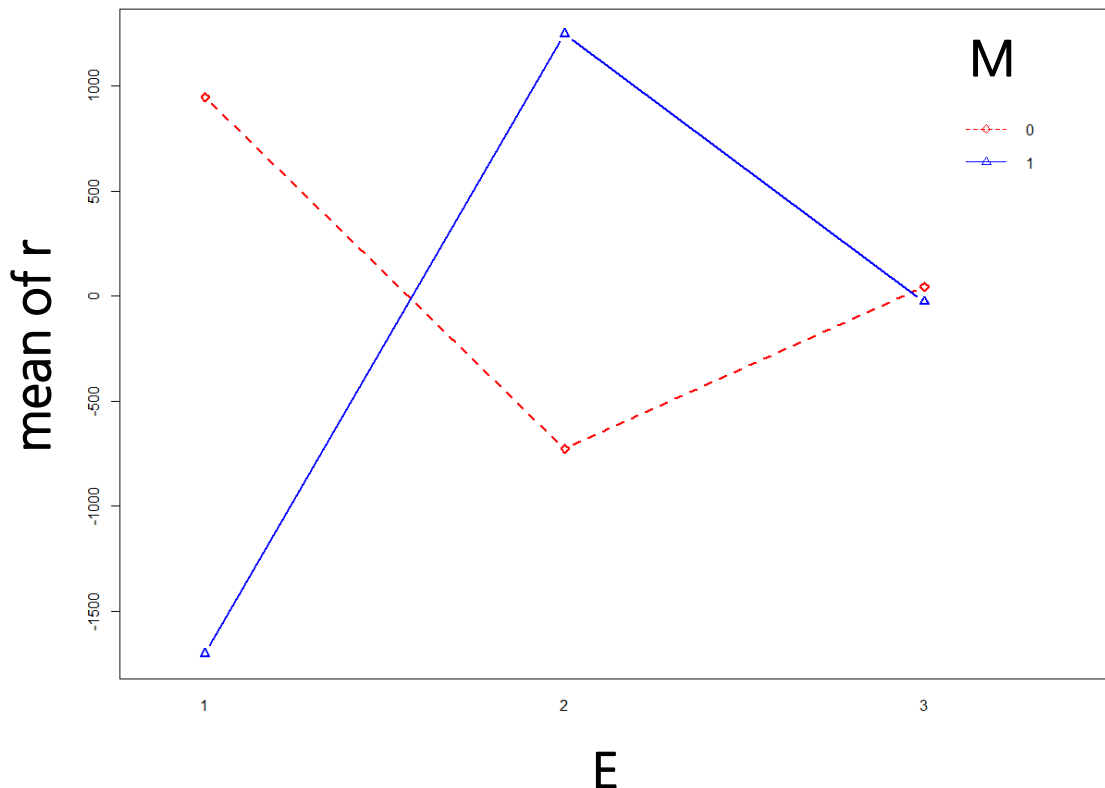




# 상호작용 변수가 필요할까?



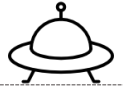
```
r = resid(salary.lm)
interaction.plot(salary.table$E, salary.table$M, r,
                 type='b', col=c('red','blue'),
                 lwd=2, pch=c(23,24))
```



더 확실히 하기 위해서  
교육수준에 따른 평균 잔차의  
분포를 관리직 유무에 따라서  
그래프를 그려봤다.

두 그래프는 다른 양상을 보이는  
것으로 확인된다.

# 모형 가정하기



교육수준과 관리직 유무가  
상호작용을 보이는지 알고 싶다.

$$S = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot E + \beta_3 \cdot M + \beta_4 \cdot \text{interactions} + \varepsilon_i$$

# 상호작용 변수의 적합



$$S = \beta_0 + \beta_1 X + \beta_2 E_2 + \beta_3 E_3 + \beta_4 M + \beta_5 E_2 M + \beta_6 E_3 M + \varepsilon$$

```
model_EM = lm(S ~ X + M + E + E:M, salary.table)
summary(model_EM)
```

```
Call:
lm(formula = S ~ X + M + E + E:M, data = salary.table)

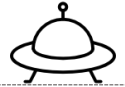
Residuals:
    Min       1Q   Median       3Q      Max
-928.13  -46.21   24.33   65.88  204.89

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9472.685     80.344  117.90  <2e-16 ***
X             496.987      5.566   89.28  <2e-16 ***
M1           3981.377     101.175   39.35  <2e-16 ***
E2           1381.671      77.319   17.87  <2e-16 ***
E3           1730.748     105.334   16.43  <2e-16 ***
M1:E2         4902.523     131.359   37.32  <2e-16 ***
M1:E3         3066.035     149.330   20.53  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 173.8 on 39 degrees of freedom
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9986
F-statistic: 5517 on 6 and 39 DF,  p-value: < 2.2e-16
```

혹은, `lm(S ~ X + E*M, salary.table)` 이라고 해도 같은 결과를 반환해줌^^

# 모델간 F 검정



보통 2가지 모델을 비교할 때는 F test를 진행한다.  
ANOVA 는 두개 이상의 집단에 F test를 진행할 때 쓴다.

귀무가설  $H_0 : \beta_5 = \beta_6 = 0$

```
anova(salary.lm, model_EM)
```

```
> anova(salary.lm, model_EM)
Analysis of Variance Table
```

```
Model 1: S ~ E + M + X
```

```
Model 2: S ~ X + M + E + E:M
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	41	43280719				
2	39	1178168	2	42102552	696.84	< 2.2e-16 ***

```
---
```

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

여기서 p-value가 0에 가까우므로, 어떤 유의수준을 쓰더라도 model2를 적합하는 것을 채택한다. 따라서 E와 M 사이에 상호작용이 존재한다고 볼 수 있다.

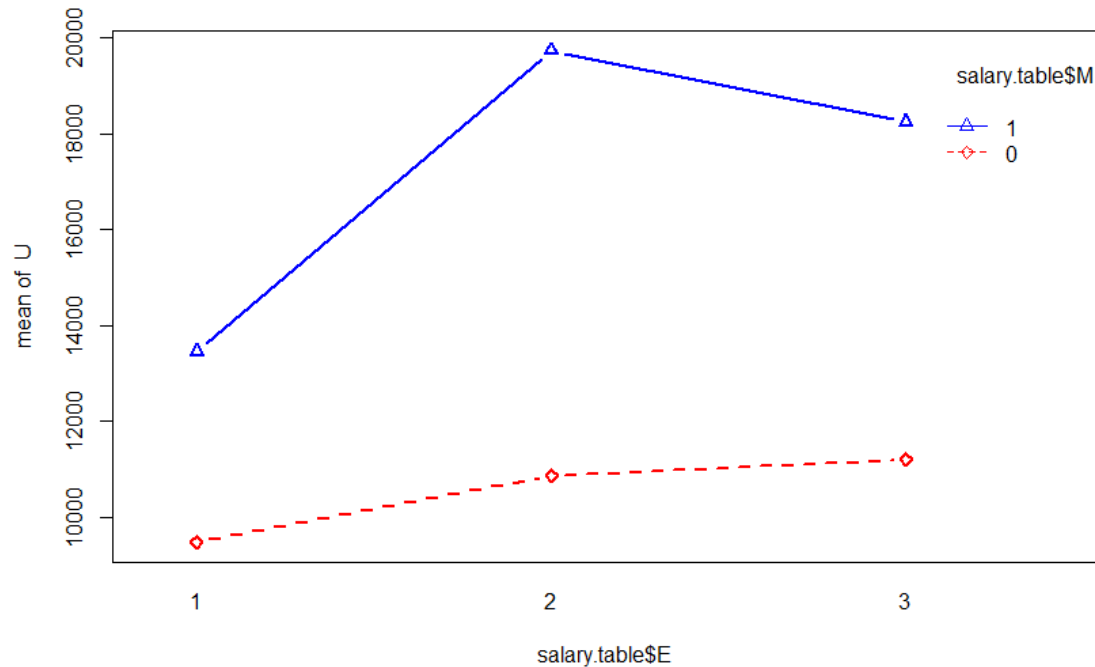
# 시각적 판단



```
U = salary.table$S - salary.table$X * model_EM$coef['X']  
interaction.plot(salary.table$E, salary.table$M, U, type='b', col=c('red','blue'), lwd=2, pch=c(23,24))
```

U = S에서 X에 대한 효과를 모두 제거한 값

```
interaction.plot(salary.table$E, salary.table$M, U, type='b', col=c('red','blue'), lwd=2, pch=c(23,24))
```



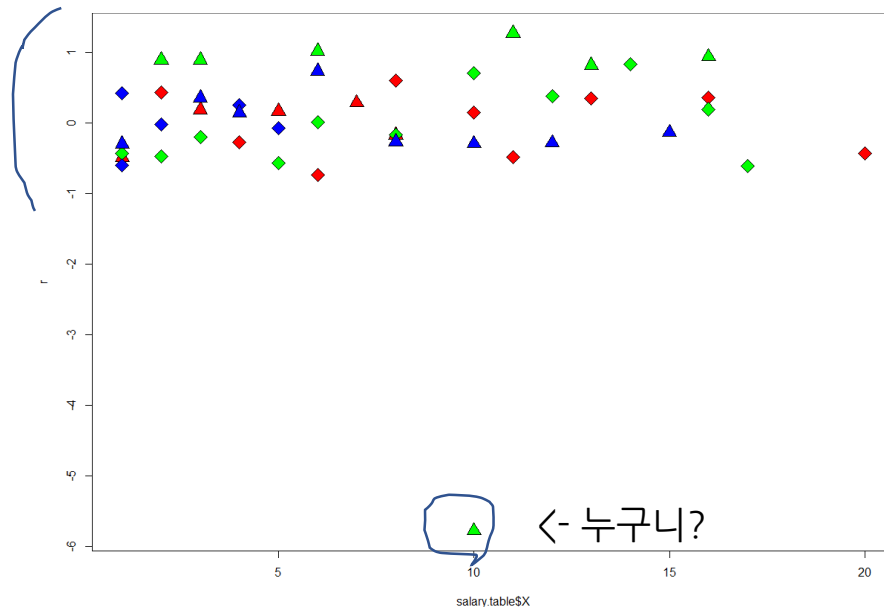
우리가 잘 검정했다는 것을 알 수 있다.

# 이상치 제거



그럼 왜 상호작용 모델에서 잔차가 표준정규분포를 따르지 않을까?  
혹시 이상치가 있는지 살펴보도록 하자.

```
r = rstandard(model_EM)
plot(salary.table$X, r, type='n')
for (i in 1:3) {
  for (j in 0:1) {
    subset <- as.logical((salary.table$E == i) * (salary.table$M == j))
    points(salary.table$X[subset], r[subset], pch=symbols[j+1], bg=colors[i], cex=2)
  }
}
```



# 이상치 제거



그래프에서 발견한 이상치를 제거하기 위해서 정확히 알아보자.  
'car' 패키지의 outlierTest 함수는 이상치가 어디에서 존재하는지 검사해준다

```
library(car)  
outlierTest(model_EM)
```

```
> outlierTest(model_EM)  
      rstudent unadjusted p-value Bonferonni p  
33  -14.95083      1.6769e-17      7.714e-16
```

33번째 관측치가 이상치라는 것을 보여준다.  
표 해석은 진단파트에서 다룹니다 ^^! 여기선 그냥 outlier 유무만 !

# 이상치 제거



33번째 이상치를 정상적으로 제거해주고 다시 모델을 적합해보자.

```
subs33 = c(1:length(salary.table$S))[-33]
salary.lm33 = lm(S ~ E + X + M, data=salary.table, subset=subs33)
model_EM33 = lm(S ~ E + X + E:M + M, data=salary.table, subset=subs33)
summary(model_EM33)
```

```
Call:
lm(formula = S ~ E + X + E:M + M, data = salary.table, subset = subs33)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-112.884	-43.636	-5.036	46.622	128.480

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	9458.378	31.041	304.71	<2e-16 ***
E2	1384.294	29.858	46.36	<2e-16 ***
E3	1741.336	40.683	42.80	<2e-16 ***
X	498.418	2.152	231.64	<2e-16 ***
M1	3988.817	39.073	102.08	<2e-16 ***
E2:M1	5049.294	51.668	97.73	<2e-16 ***
E3:M1	3051.763	57.674	52.91	<2e-16 ***

---

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

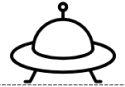
Residual standard error: 67.12 on 38 degrees of freedom

Multiple R-squared: 0.9998, Adjusted R-squared: 0.9998

F-statistic: 3.543e+04 on 6 and 38 DF, p-value: < 2.2e-16



# 또.. 다시.. F TEST..



33번째 이상치를 정상적으로 제거해주고 다시 모델을 적합해보자.

귀무가설  $H_0$  : E와 M 사이에 상호작용이 존재하지 않는다.

```
> anova(salary.lm33, model_EM33)
Analysis of Variance Table
```

```
Model 1: S ~ E + X + M
```

```
Model 2: S ~ E + X + E:M + M
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	40	43209096				
2	38	171188	2	43037908	4776.7	< 2.2e-16 ***

```
---
```

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

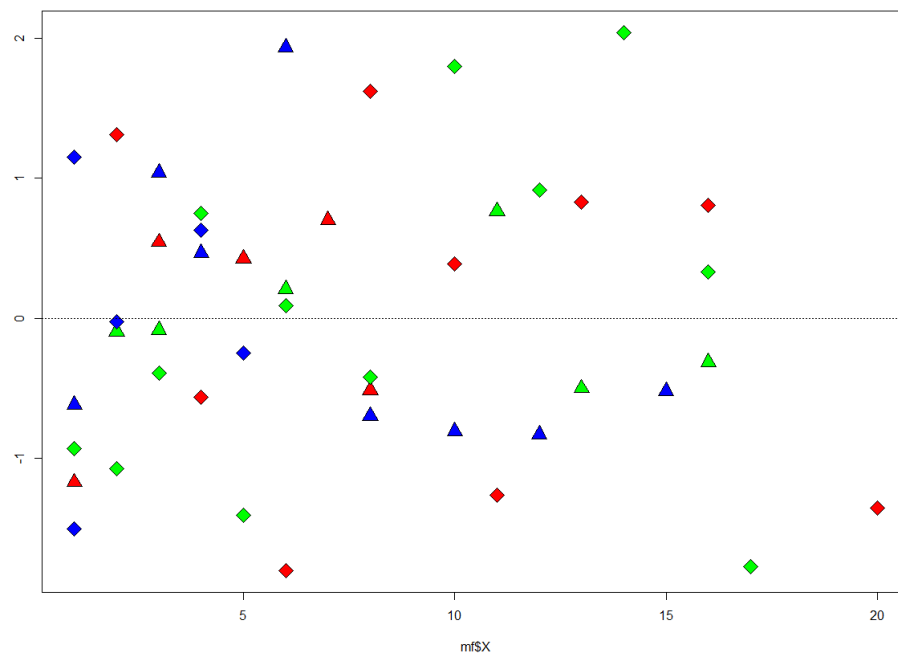
여기서 p-value값이 0에 가까우므로, 어떤 유의수준을 쓰더라도 귀무가설을 기각하고 model\_EM을 적합하는 것을 채택한다. 따라서 E와 M 사이에 상호작용이 존재한다고 볼 수 있다.

# 또.. 잔차.. 플롯.. 확인..



```
r = rstandard(model_EM33)
mf = model.frame(model_EM33)
plot(mf$X, r, type='n')
for (i in 1:3) {
  for (j in 0:1) {
    subset <- as.logical((mf$E == i) * (mf$M == j))
    points(mf$X[subset], r[subset], pch=symbols[j+1], bg=colors[i], cex=2)
  }
}
abline(h=0, lty=3)
```

예쁘다.

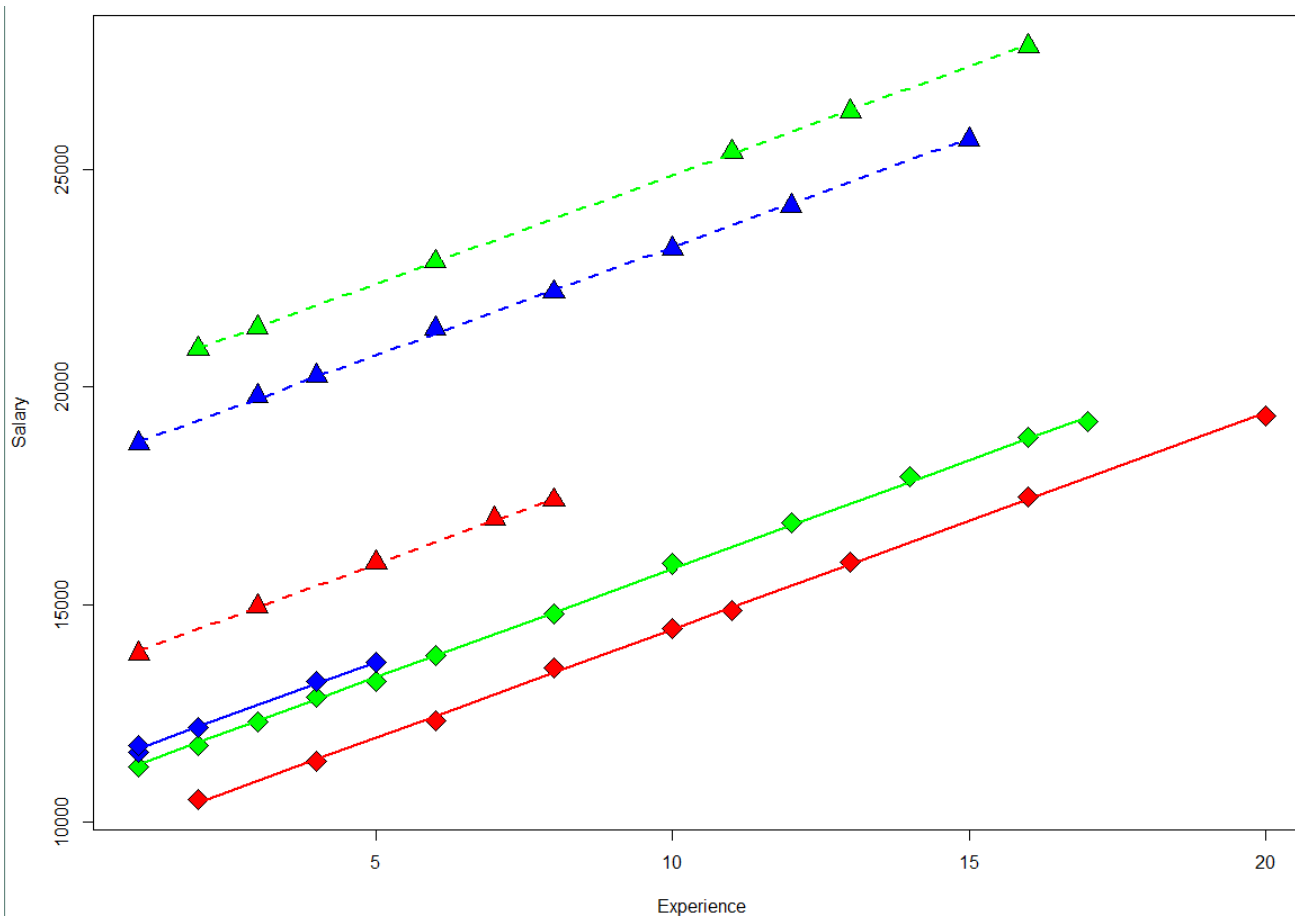


# 그래프로 회귀직선 그려보기



```
plot(mf$X, mf$S, type='n', xlab='Experience', ylab='Salary')
colors <- c('red', 'green', 'blue')
ltys <- c(2,3)
symbols <- c(23,24)
for (i in 1:3) {
  for (j in 0:1) {
    subset <- as.logical((mf$E == i) * (mf$M == j))
    points(mf$X[subset], mf$S[subset], pch=symbols[j+1], bg=colors[i], cex=2)
    lines(mf$X[subset], fitted(model_EM33)[subset], lwd=2, lty=ltys[j], col=colors[i])
  }
}
```

# 그래프로 회귀직선 그려보기



깔끔!

# 여기서 잠깐 !



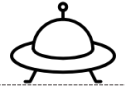
만약에 interaction term 이 있는데 main effect를  
테스트하려고 하면 어떻게 될까 ?

H1 : S에 대한 M의 효과가 없다.

```
model_main = lm(S ~ X+E+E:M, salary.table)
```

```
anova(model_EM, model_main)
```

# 여기서 잠깐 !



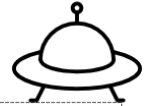
만약에 interaction term 이 있는데 main effect를  
테스트하려고 하면 어떻게 될까 ?

H1 : S에 대한 M의 효과가 없다.

```
> anova(model_EM, model_main)
Analysis of Variance Table

Model 1: S ~ X + M + E + E:M
Model 2: S ~ X + E + E:M
  Res.Df    RSS Df Sum of Sq  F Pr(>F)
1      39 1178168
2      39 1178168  0 1.8626e-09
```

아예 F값이 뜨지 않음. 즉 interaction term은 상호작용에 대해서 테스트를 할 때만  
쓴다. 다르게 말해서, 우리는 우리가 원하는 가설에 따라 다른 모델을 적절히  
적합할 필요가 있다.



**질문?**

**저도 잘 모릅니다.**