회귀모형

사기 1조

김연모 김재훈 신은아 장은조

목차

단순회귀분석의 개념

장은조

회귀분석이란?

기본적으로 하나 이상의 독립변인(들)이 한 단위 변할 때,
종속변인이 얼마나 변할 것인지, 다시 말해 하나 이상의
독립변인(들)이 종속변인에 미치는 영향력을 예측하는 데 주로 사용하는 **통계분석기법**이다.

회귀라는 용어의 기원은 Galton이 자녀의 키와 부모의 키의 관계를 분석한 논문에서 찾을 수 있다. 그는 928명의 성인 자녀와 그들의 부모의 키 사이에 직선의 관계가 있음을 발견하였고, "자녀의 키는 부모의 키가 그면 대체적으로 크나 부모의 키보다는 작으며 전체 자녀들의 평균키에 근접하는 경향이 있다." 라는 사실을 발표하였는데 이를 "regression toward mediocrity" 란 용어를 써서 발표하였고, 이러한 이유로 회귀라는 용어의 부적절성에도 불구하고 변수와 변수와의 관계를 도출하고자 하는 기법을 통계학에서는 회귀분석이라고 한다.

두 변수간의 함수관계를 찾는 첫 단계는 산점도

```r
year <- 1971:1982
Y <- c(5.6,3.2,4.5,4.2,5.2,2.7,4.8,4.9,4.7,4.1,4.4,5.4)
X <- c(116.37,82.77,110.68,97.5,115.88,80.19,125.24,116.15,117.36,93.31,107.46,122.3)
data <- data.frame(year,Y,X)
data
plot(data$X,data$Y,xlab="열매개수",ylab="수확량",main="수확량과 열매개수의 산점도 ",col="red")
```

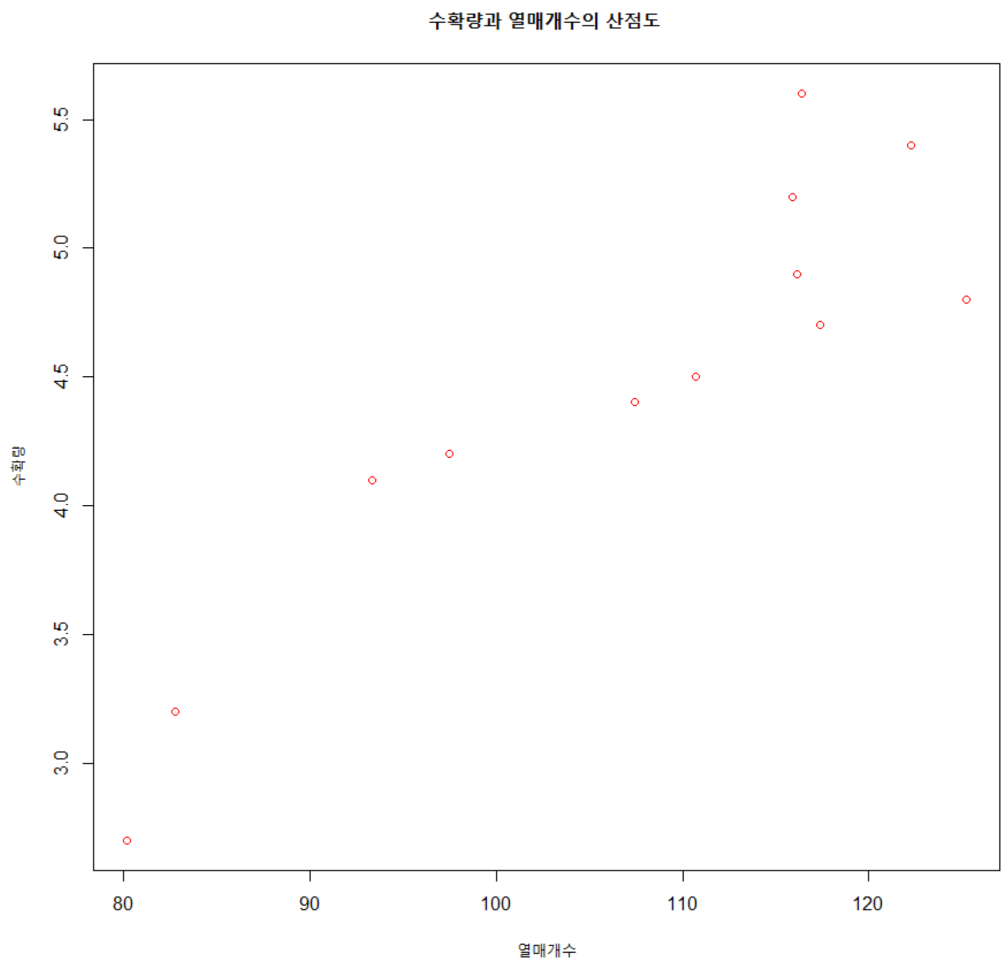
수확량은 열매개수에 비례한다는 것을 알 수 있다.
이러한 사실을 감안하면 다음의 관계식을 생각할 수 있다.

$$\boxed{Y = \beta_0 + \beta_1 X}$$

수확량 측정오차, 수확량에 영향을 미치는 다른 변수들
(토양의 비옥도, 비료의 살포량, 기후 등등)

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

ε 오차 (통제할 수 없는 부분)



회귀분석 모델

$Y = \boxed{f(X)} + \varepsilon$ 설명변수가 1개인 회귀모형

회귀함수 f(x)로 고려되는 가장 간단한 함수는 **선형함수** 이다.

$$\boxed{f(x) = \beta_0 + \beta_1 x}$$

X가 Y에 미치는 영향이 선형적임을 의미한다.

오차의 구성요소

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

n개의 자료를 반영

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

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

오차

- 고정요소 　가정하는 모형이 두 변수 사이에 존재하는 참의 관계식을 반영하지 못할 때 발생
- 확률적 요소
  - 측정오차 　100%의 정확도를 가지고 변수의 값을 측정한다는 것은 불가능
  - 모형에 포함되어야 하는 설명변수가 제외됨으로써 나타날 수 있는 오차
  - 자연발생적으로 생겨나 임의적으로 통제할 수 없는 순수오차

회귀모형 구성요소

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

$$i = 1,2, \dots, n$$

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

오차항의 가정 (뒤에서 자세히 언급)

$\beta_0,\ \beta_1$: 모수로서 회귀계수

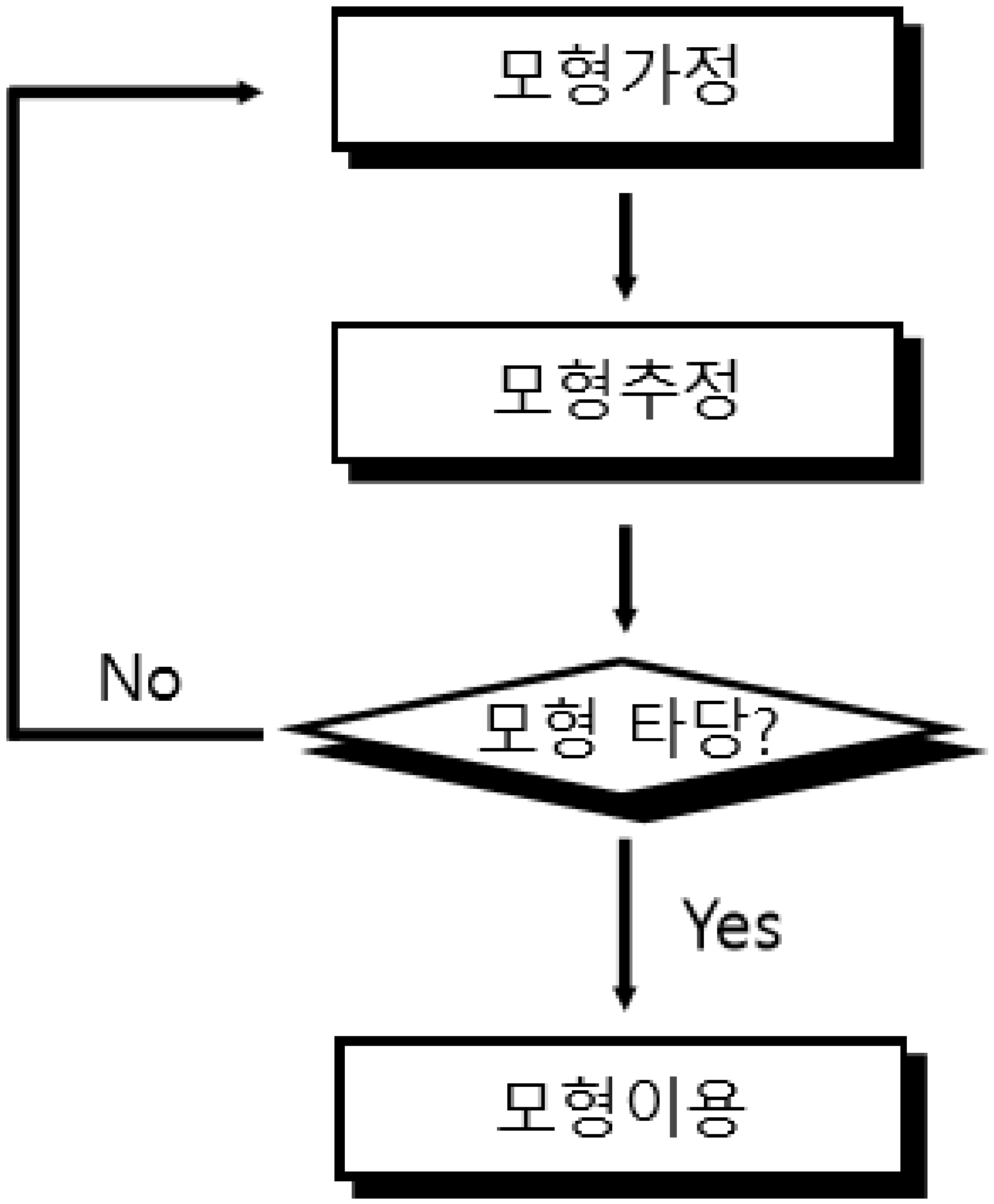
ε_i : 통계적 오차 - > 확률적인 요소

x_i : 설명변수 > 상수로 주어짐

$\beta_0 + \beta_1 x_i$: 상수 -> 확정적인 요소

y_i : 반응변수 > 확률변수

회귀분석의 단계

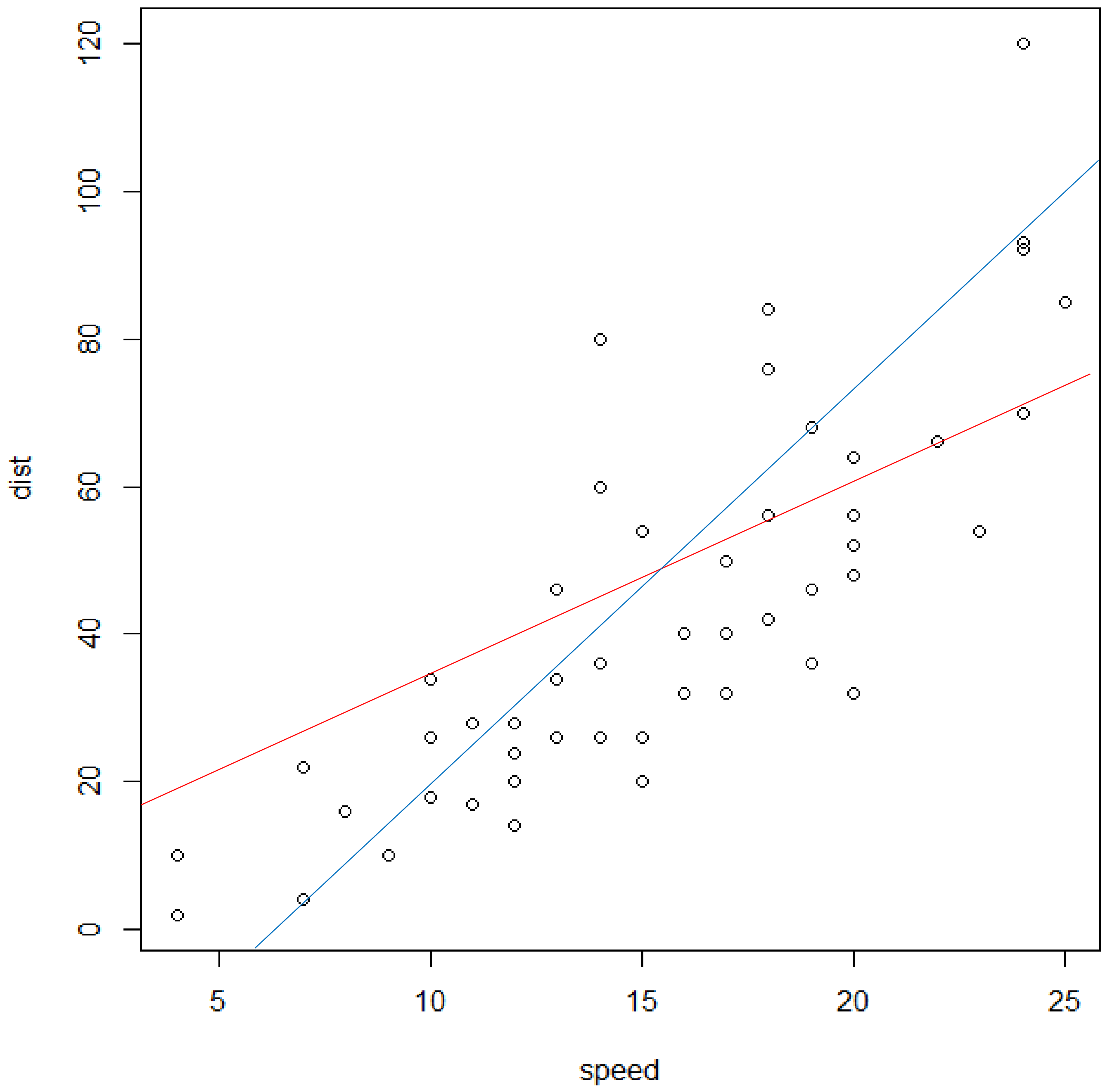


모형 추정

다음의 데이터를 더 잘 설명하는 직선은?

파란 직선?

빨간 직선?



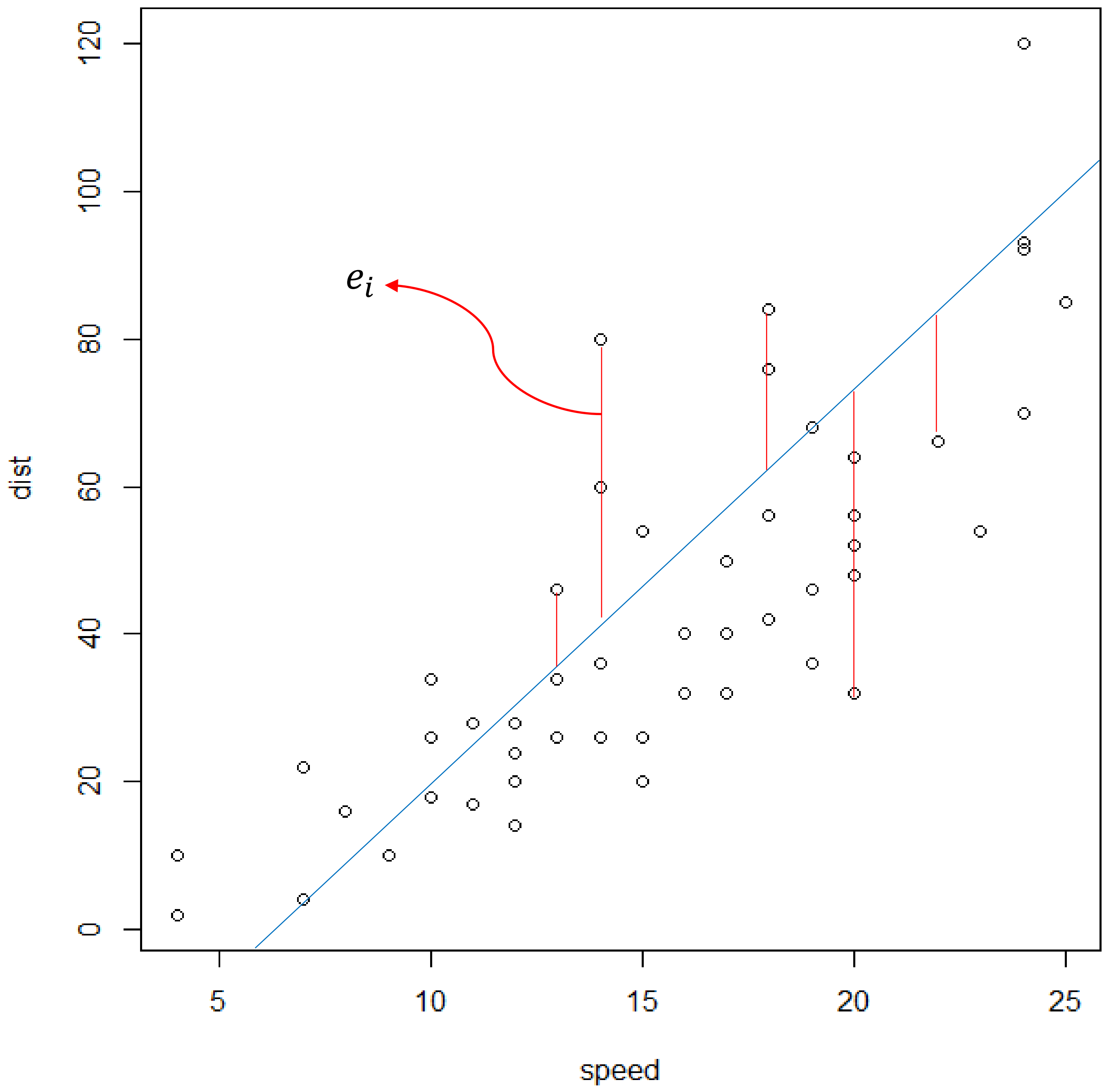
최소제곱법

각각의 관측값들과 직선 사이에는 수직거리가 발생하는데 이를 잔차 (e_i) 라고 한다

잔차는 양의값, 음의값 모두 가질 수 있다.

잔차의 제곱합이 최소값을 가질 수 있도록 모수를 추정하는 방법

즉, 내가 추정한 회귀 직선과 실제 관측값 사이의 거리의 제곱을 최소로 하는 회귀계수 추정하는 방법



최소제곱법 유도과정

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ , \quad i = 1,2,\ldots\ldots,n$$

$$Q(\hat\beta_0, \hat\beta_1) = \boxed{\sum (e_i)^2 = \sum (y_i - \hat y_i)^2 = \sum (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2}$$

SSE

$$\left.\begin{cases} \dfrac{\partial Q}{\partial \hat\beta_0}\Big|_{\hat\beta_0,\hat\beta_1} = -2\sum(y_i - \hat\beta_0 - \hat\beta_1 x_i) \overset{set}{=} 0 \\ \dfrac{\partial Q}{\partial \hat\beta_1}\Big|_{\hat\beta_0,\hat\beta_1} = -2\sum(y_i - \hat\beta_0 - \hat\beta_1 x_i)x_i \overset{set}{=} 0 \end{cases}\right\} \text{Normal equation}$$

$$\begin{cases} \sum y_i = n\hat\beta_0 + \hat\beta_1 \sum x_i \\ \sum y_i x_i = \hat\beta_0 \sum x_i + \hat\beta_1 \sum x_i^2 \end{cases}$$

$$\boxed{\hat\beta_0} = \bar y - \hat\beta_1 \bar x$$

$$\boxed{\hat\beta_1} = \frac{S_{xy}}{S_{xx}}$$

최소제곱추정량(LSE)

질문

"최소제곱법에 의해 추정한 회귀식이 가장 적절하다고 할 수 있나?"

"최소제곱법은 항상 쓸 수 있는건가?"

가우스- 마코브 정리

1. 오차변수의 기대값은 0이다
2. 오차변수와 독립변수의 공분산은 0 이다
3. 오차변수의 분산은 일정한 상수이다
4. 오차변수들 사이의 공분산은 0이다
5. 오차변수는 정규분포를 따른다.

1~4의 조건을 만족한다면 최소제곱법은 선형추정치 중 가장 좋은 불편추정량이 되며,
5번 조건까지 만족하게 된다면 선형추정치 중 가장 좋은 불편추정량 이면서 분산까지 가장 작은 추정량이 된다.

즉, 선형관계의 척도에 대해 최소제곱법이 오차변수의 4~5가지 조건을 만족한다면
가장 좋은 선형관계를 보여주는 방법이라는 것을 보여주는 것이 가우스-마코브 정리이다.

변동분해 - 분산분석

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$SSE = \Sigma e_i^2 = \sum (y_i - \hat{y}_i)^2$$

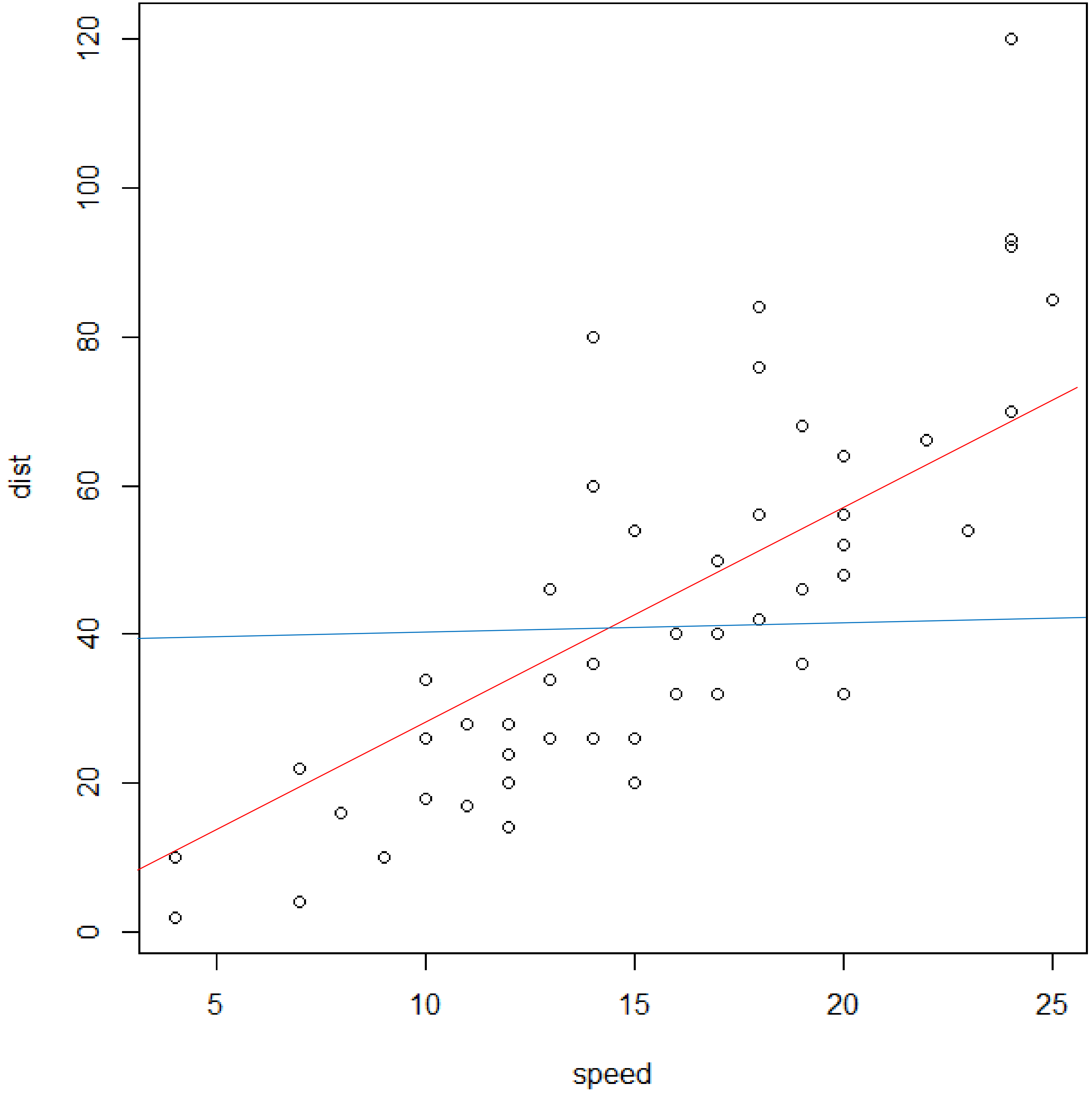
>> 기울기가 의미 있으며 변수 사이에는 선형의 관계가 존재한다

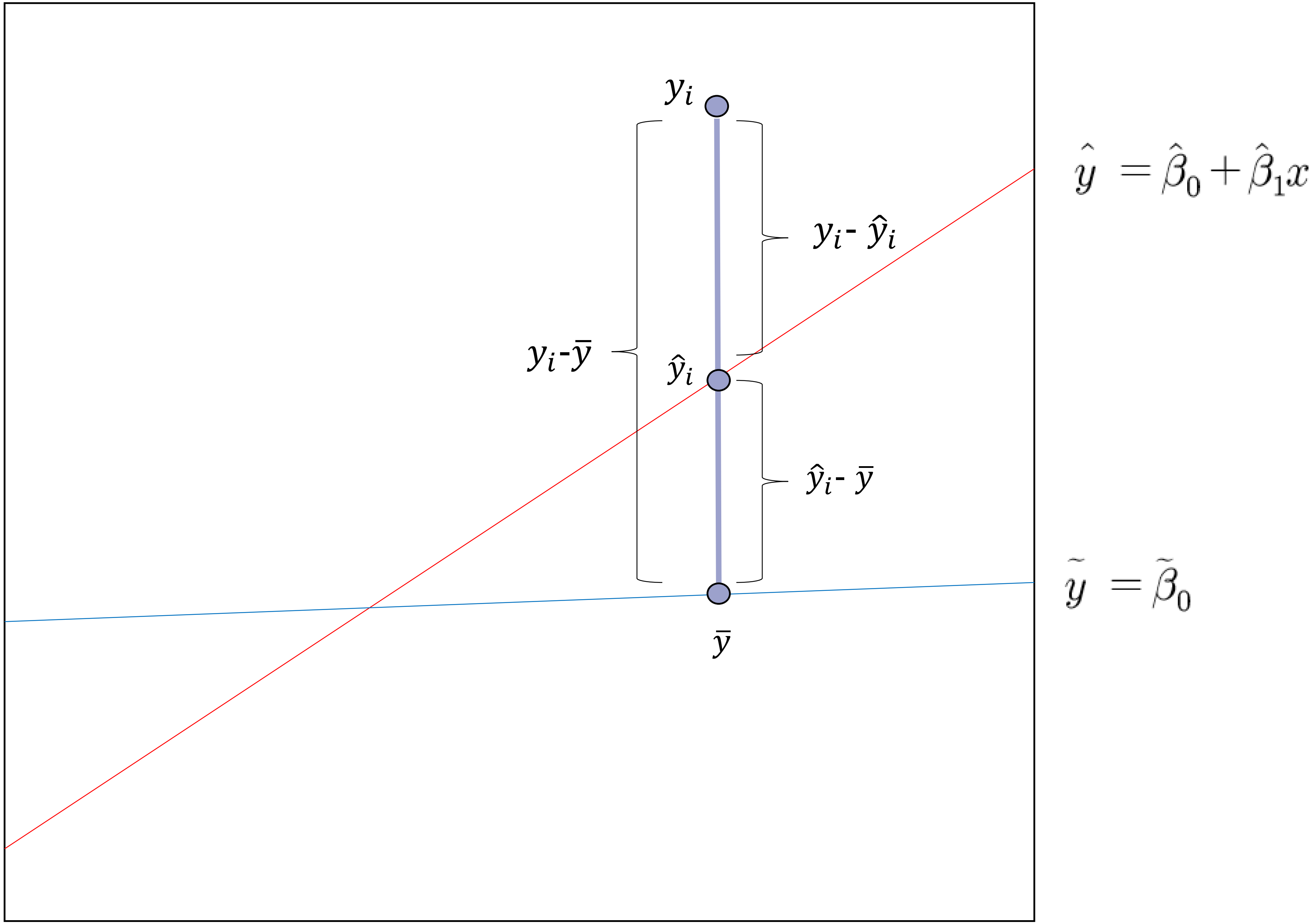
$$y_i = \beta_0 + \epsilon_i$$

$$\tilde{y}_i = \tilde{\beta}_0$$

$$SSE = \Sigma e_i^2 = \sum (y_i - \bar{y})^2$$

>> 반응변수가 설명변수의 영향을 받지 않는다




y_i
$y_i - \hat{y}_i$
$y_i - \bar{y}$
$\hat{y}_i$
$\hat{y}_i - \bar{y}$
$\bar{y}$
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\tilde{y} = \tilde{\beta}_0$

변동분해

$$\underset{\text{(총)편차}}{y_i - \bar{y}} = \boxed{\underset{\text{잔차}}{(y_i - \hat{y}_i)}} \boxed{+ \underset{\text{추측값의 편차}}{(\hat{y}_i - \bar{y})}}$$

회귀식으로 설명 불가능한 편차 회귀식으로 설명 가능한 편차

$$\boxed{\underset{\text{총제곱합}}{\sum (y_i - \bar{y})^2}} = \boxed{\underset{\text{잔차제곱합}}{\sum (y_i - \hat{y}_i)^2}} + \boxed{\underset{\text{회귀제곱합}}{\sum (\hat{y}_i - \bar{y})^2}} \boxed{\cancel{+2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}}^{\,0}$$

SST SSE SSR

분산분석표 (ANOVA)

어떤 통계량S의 자유도란 S를 구성하고 있는 기본요소 중 서로 독립인 기본요소의 개수

Source of Variation	Degree of Freedom (DF)	Sum of Squares(SS)	Mean Square(MS)	F
Regression	1	SSR	MSR	MSR/MSE
Error	n-2	SSE	MSE	
Total	n-1	SST		

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ = 총 편차

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0$$

1개의 제약조건으로
통계량 SST를 구성하고 있는 n개의 구성요소 중
n-1개의 요소가 독립적이라고 할 수 있다.

잔차의 합은 0

분산분석표 (ANOVA)

Source of Variation	Degree of Freedom (DF)	Sum of Squares(SS)	Mean Square(MS)	F
Regression	1	SSR	MSR	MSR/MSE
Error	n-2	SSE	MSE	
Total	n-1	SST		

$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ = 회귀식으로 설명 불가능한 편차

$$Q(\hat\beta_0,\hat\beta_1) = \sum(e_i)^2 = \sum(y_i-\hat y_i)^2 = \sum(y_i-\hat\beta_0-\hat\beta_1 x_i)^2$$

$$\left\{\begin{aligned} \frac{\partial Q}{\partial \hat\beta_0}\Big|_{\hat\beta_0,\hat\beta_1} &= -2\sum(y_i-\hat\beta_0-\hat\beta_1 x_i) \overset{set}{=} 0 \\ \frac{\partial Q}{\partial \hat\beta_1}\Big|_{\hat\beta_0,\hat\beta_1} &= -2\sum(y_i-\hat\beta_0-\hat\beta_1 x_i)x_i \overset{set}{=} 0 \end{aligned}\right\}$$

2개의 제약조건으로
통계량 SSE를 구성하고 있는 n개의 구성요소 중
n-2개의 요소가 독립적이라고 할 수 있다.

분산분석표 (ANOVA)

Source of Variation	Degree of Freedom (DF)	Sum of Squares(SS)	Mean Square(MS)	F
Regression	1	SSR	MSR	MSR/MSE
Error	n-2	SSE	MSE	
Total	n-1	SST		

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

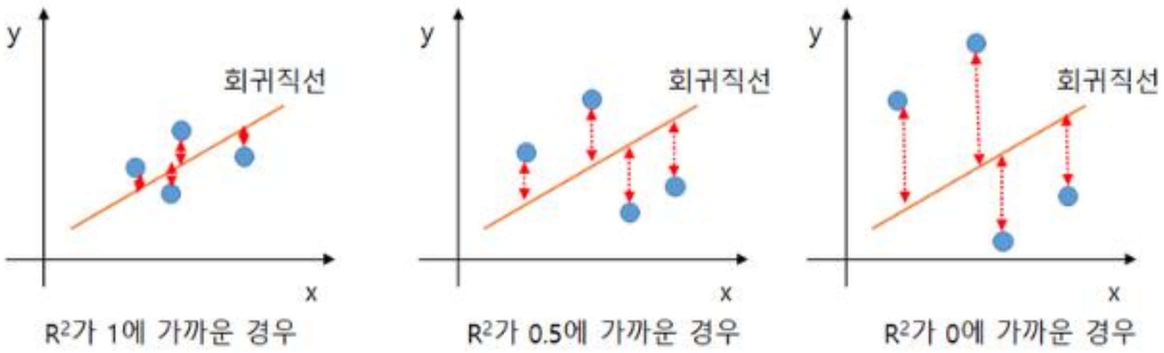
- y의 총변동 중에서 회귀모형에 의해 설명이 되는 변동의 크기
- 결정계수의 범위는 0 과 1사이 (0,1포함)
- 결정계수가 커질수록 회귀에 대한 설명력이 커짐

실제 데이터가 회귀직선에 매우 밀접하게 분포

질문

“결정계수는 이제 알겠는데 상관계수랑은 다른건가요?” YES

결정계수



상관계수

$$R = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\mathrm{var}(y)}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

X와 Y간의 선형적인 관계를 나타내는 척도

상관계수의 범위는 -1 ~ 1 (-1,1포함)

1에 가까울 수록 양의 상관관계, -1에 가까울수록 음의 상관관계를 의미

0에 가까울수록 두 변수 간에 선형적인 관계가 없다고 볼 수 있음

분산분석표 (ANOVA)

Source of Variation	Degree of Freedom (DF)	Sum of Squares(SS)	Mean Square(MS)	F
Regression	1	SSR	MSR	MSR/MSE
Error	n-2	SSE	MSE	
Total	n-1	SST		

$H_0: \ y_i = \beta_0 + \epsilon_i,$

$H_1: \ y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

⇔ $H_0: \ \beta_1 = 0$

$H_1: \ \beta_1 \neq 0$

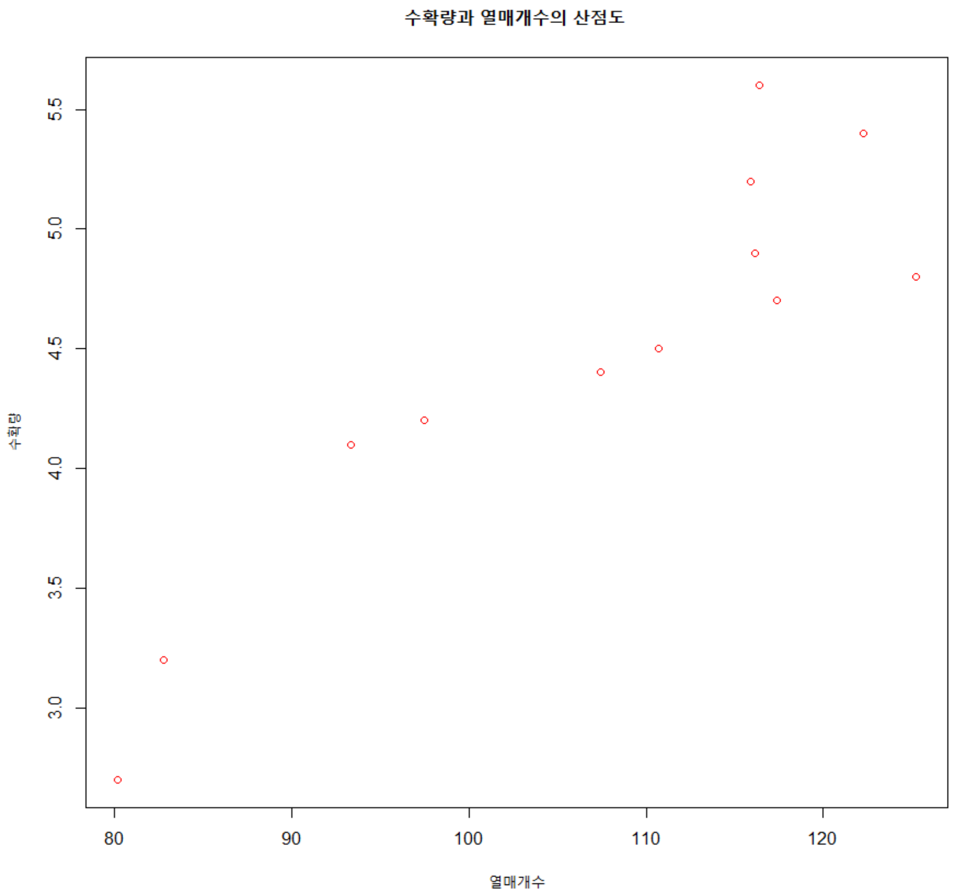
검정통계량 $F = \dfrac{MSR}{MSE} \sim F_{(1,\ n-2)}$ under H_0

$F > F_{\alpha,(1,\ n-2)}$(임계값)이면, 유의수준 α에서 H_0을 기각

(또는 $\alpha \geq$ p-value(유의확률)이면, H_0을 기각)

F-검정 (적합된 회귀모형 $(y_i = \beta_0 + \beta_1 x_i + \varepsilon_i)$의 유의성 검정)

예제를 통해 알아보아요




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

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

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

$$\boxed{\hat{y} = -17.579 + 3.932x}$$

예제를 통해 알아보아요

```
Analysis of Variance Table

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

$H_0 : \ y_i = \beta_0 + \epsilon_i,$

$H_1 : \ y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$\Leftrightarrow \quad H_0 : \ \beta_1 = 0$

$\quad H_1 : \ \beta_1 \neq 0$

검정통계량 $F = \dfrac{MSR}{MSE} \sim F_{(1,\ n-2)}$ under H_0

$F > F_{\alpha,(1,\ n-2)}$(임계값)이면, 유의수준 α에서 H_0을 기각

(또는 $\alpha \geq$ p-value(유의확률)이면, H_0을 기각)

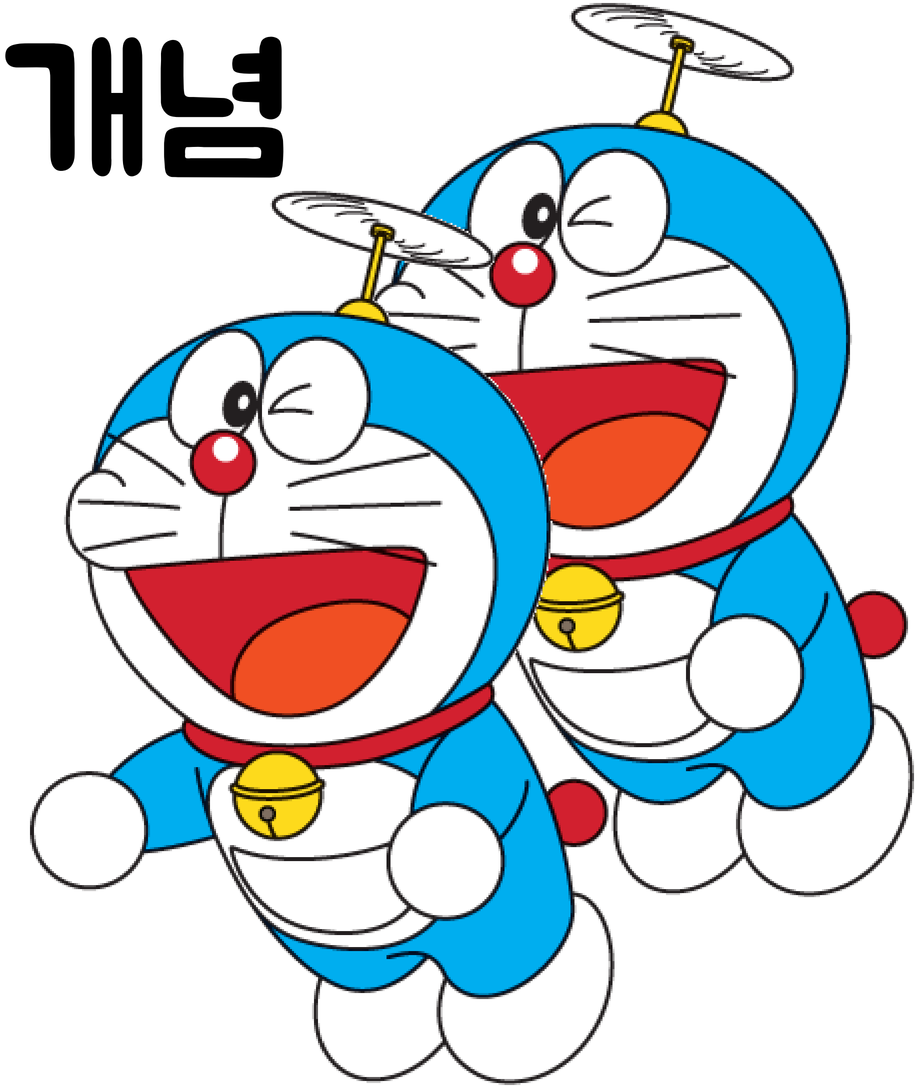
결론 : H0을 기각 , $\hat{y} = -17.579 + 3.932x$ **채택**

10분 쉬었다 가죠~

"ASK GO TH THE BLUE"

다중회귀분석의 개념

신은아



통계적 추론

단순회귀 모형(OLS Model)

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

Q. 여기에서 y_i의 분포는 어떻게 될까?!

x_i : 상수, y_i : 확률변수이고,
정규분포에 상수를 더해도 정규분포를 따른다!

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

$$\begin{aligned} Var(y_i) &= Var(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= Var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

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

$\hat{\beta}_0$과 $\hat{\beta}_1$의 분포

$$E\widehat{\beta}_0 = \beta_0$$
$$E\widehat{\beta}_1 = \beta_1$$
$$Var(\widehat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]$$
$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2} = \frac{\sigma^2}{S_{xx}}$$

$\hat{\beta}_0$과 $\hat{\beta}_1$은 모두 y의 선형함수이므로
→정규분포를 따른다
(정규분포의 선형결합은 정규분포를 따름)

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum(x_i - \overline{x})^2}\right]\right)$$
$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \overline{x})^2}\right)$$

β_1에 대한 검정

β_1에 대한 불편추정량 b_1을 사용→ b_1의 분포를 알아야함 $\boldsymbol{b_1 \sim N(B_1, \frac{\sigma^2}{S(xx)})}$

1. 가설의 설정

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

2. 검정통계량과 분포

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

3. 기각역

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

4. P-value $\quad \mathrm{P}[T > |t_0|]$, where T ~ $t_{(n-2)}$

β_0에 대한 검정

β_0에 대한 불편추정량 b_0을 사용→ b_0의 분포를 알아야함 $\boldsymbol{b_0 \sim N(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S(xx)}\right)\sigma^2)}$

1. 가설의 설정

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

2. 검정통계량과 분포

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

3. 기각역

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

4. P-value $\quad \mathrm{P}[T > |t_0|]$, where T ~ $t_{(n-2)}$

선형회귀분석의 기본 가정

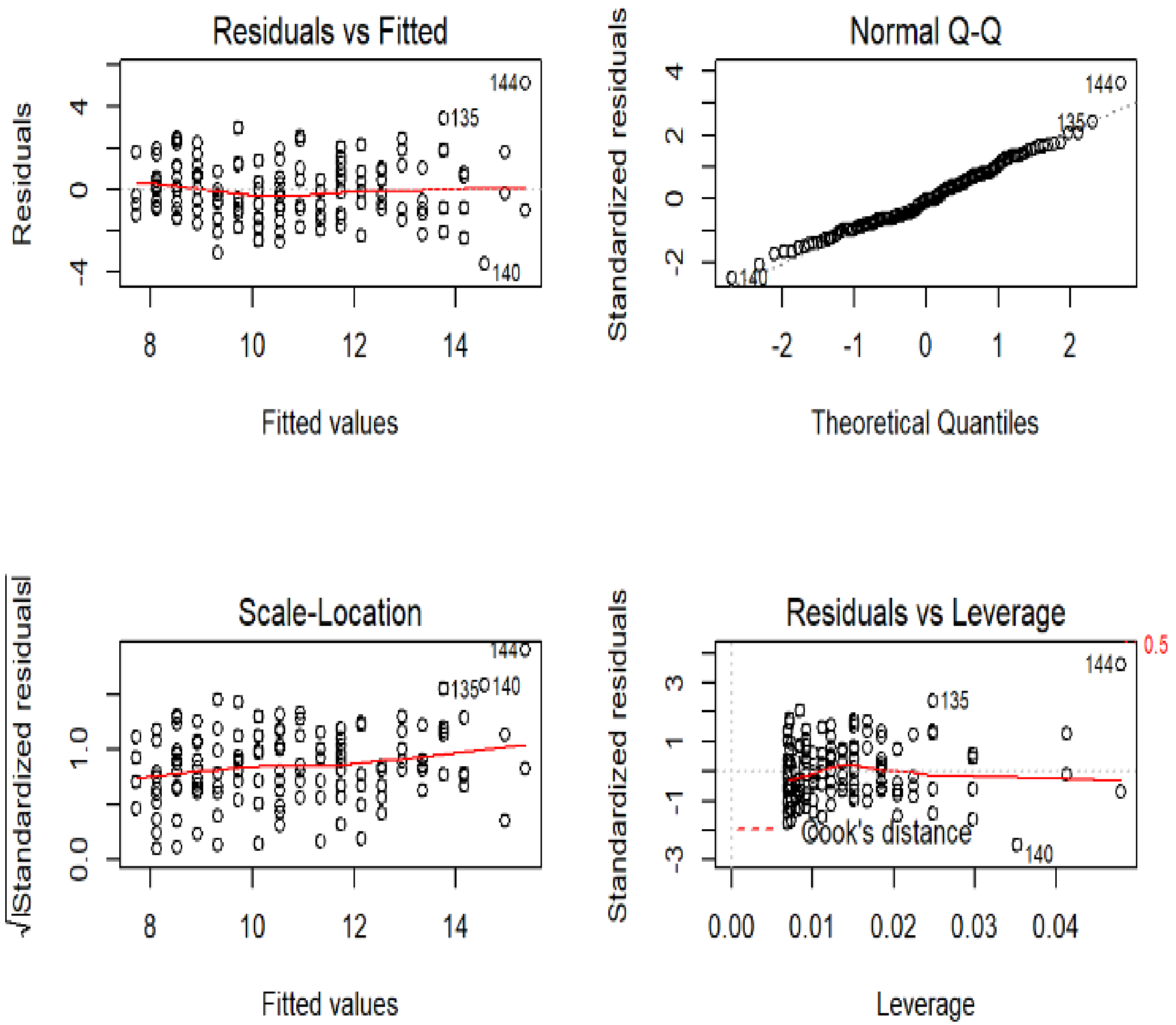
회귀분석에서는 모수들에 대한 추정을 실시한 후
→추정된 회귀선을 이용해 전제조건으로 주어진 기본 가정들에 대한 검증을 실시해야함

<선형회귀분석에서 전제조건으로 주어지는 기본 가정>

1. 선형성(linearity) : 두 변수 X와 Y의 관계는 선형관계식으로 설명할 수 있다.
2. 등분산성(homoscedasticity) : 오차항 ϵ의 분산은 모든 X값에 대해 동일하다.
3. 독립성(independence) : 오차항들은 서로 독립이다.
4. 정규성(normality) : 오차항은 정규분포를 따른다.

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

R의 plot 함수



R의 plot함수로
→등분산성, 정규성, 이상치 판단 가능

1. 회귀로 예측된 Y값에 대한 잔차 도표 → 등분산성
2. 정규 Q-Q(quantile-quantile) 도표 → 정규성
3. 척도 위치 도표 → 등분산성(이상치)
4. 잔차와 지렛대(leverage)에 대한 도표 → 이상치

어떻게 판단하는지는
뒤의 실습 파트에서 자세하게 다룰 예정!

다중선형회귀

다중선형회귀(Multiple Linear Regrssion) :
고려하는 설명변수 X의 수가 2개 이상인 선형회귀

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, \cdots, n$$

분산분석표

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

n : 관측값 수
p : 설명변수 X의 수

Adjusted R-squared

모델이 얼마나 정확한지에 대한 여부는 결정계수(R-squared)를 통해 확인 가능!

- 단순회귀에서는 **결정계수(R-squared)** ,
- 다중회귀에서는 **수정된 결정계수(Adjusted R-squared)**를 사용하여 모델의 정확성을 판단

결정계수(R-squared)	수정된 결정계수(Adjusted R-squared)
$R^2 = 1 - \frac{SSE}{SST}$	Adj $R^2 = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{MST}$

Q. 왜 수정된 결정계수를 쓰는 것일까?

설명변수의 수가 많아질수록
→ SSE의 값 ↓
→ R^2의 값이 1에 가까워짐
즉, 무조건 변수 X를 많이 적합하면 적절한 모형이라는 잘못된 결과가 나올 수 있음
→ SS-를 그 값의 자유도로 나눠줘서 값을 보정해줌

lm함수

회귀분석은 lm함수를 통해 수행할 수 있음

형태 : lm(formula, data)

formula : 회귀분석을 하기 위한 표현식
data : 회귀분석에 사용할 데이터 프레임

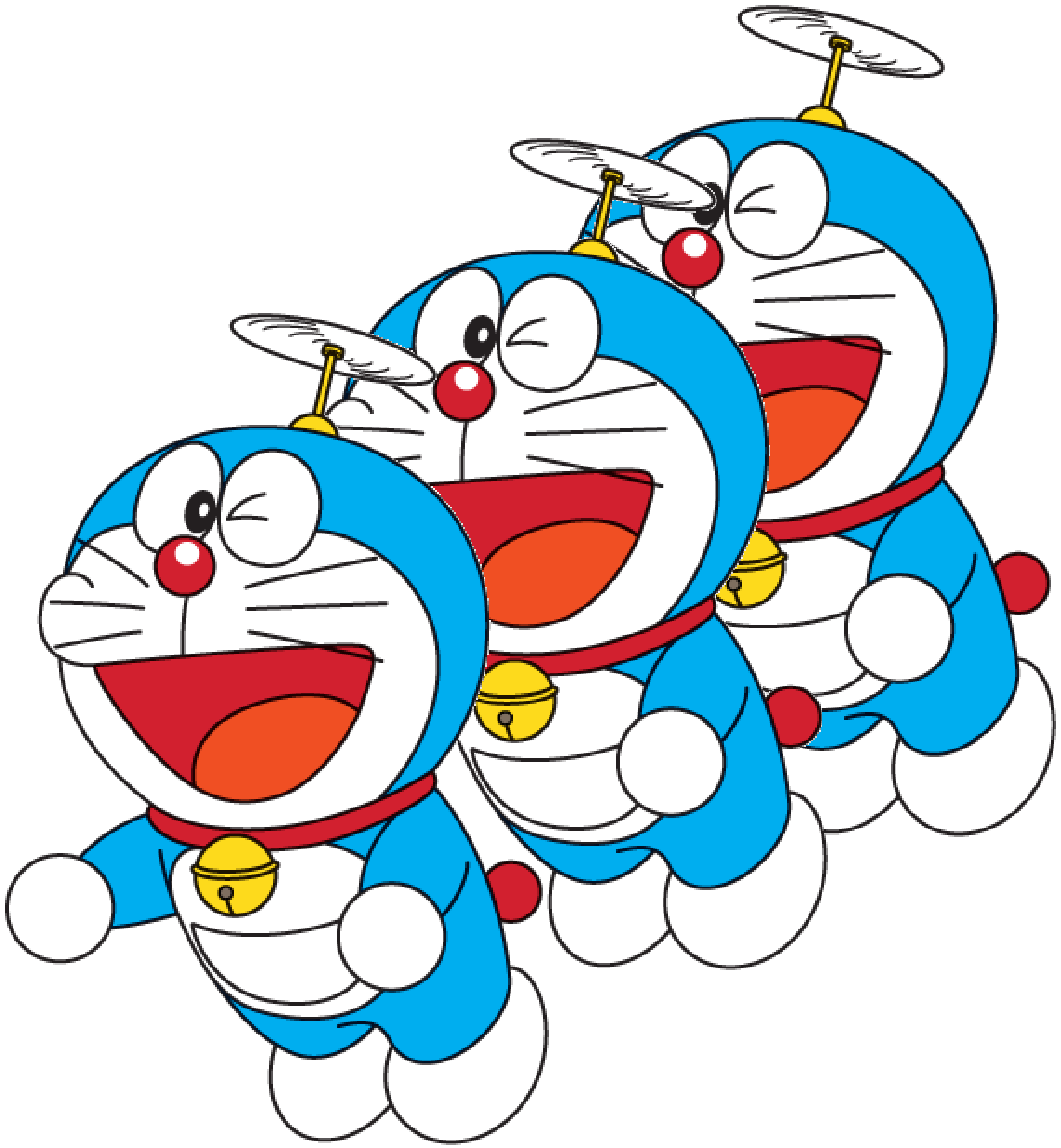
lm함수

formula

기호	의미
~	종속변수와 독립변수를 구분 짓는 기호, 종속변수는 ~ 기호의 왼쪽에 위치 Ex) 종속변수 ~ 독립변수
+	독립변수가 여러 개인 경우 + 기호로 연결 Ex) 종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3
.	전체 항목
-	선택된 독립변수 중 제외하고 싶은 독립변수는 – 기호를 입력해 삭제
:	상호작용항. "독립변수XX독립변수" 를 하나의 독립변수로 만드는 것을 표현 Ex)종속변수 ~ 독립변수A:독립변수B
*	독립변수뿐 아니라 상호관계항까지도 고려할 때 사용 Ex)" 종속변수~독립변수A*독립변수B" 는 "종속변수 ~ 독립변수A + 독립변수B + 독립변수A:독립변수B" 와 같은 의미
I()	독립변수에 특정 수식을 적용할 때 사용 Ex) "종속변수 ~ I(독립변수A*2) + 독립변수B" 인 경우 회귀분석의 독립변수는 "독립변수 A를 두 배한 값", "독립변수B" 임

회귀분석 실습

김연모



실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.1
```

여러 데이터 셋을 가지고 있는
'MASS' 패키지 불러오기

```
data(cats)
str(cats)
```

```
## 'data.frame':    144 obs. of  3 variables:
##  $ Sex: Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Bwt: num  2 2 2 2.1 2.1 2.1 2.1 2.1 2.1 2.1 ...
##  $ Hwt: num  7 7.4 9.5 7.2 7.3 7.6 8.1 8.2 8.3 8.5 ...
```

```
summary(cats)
```

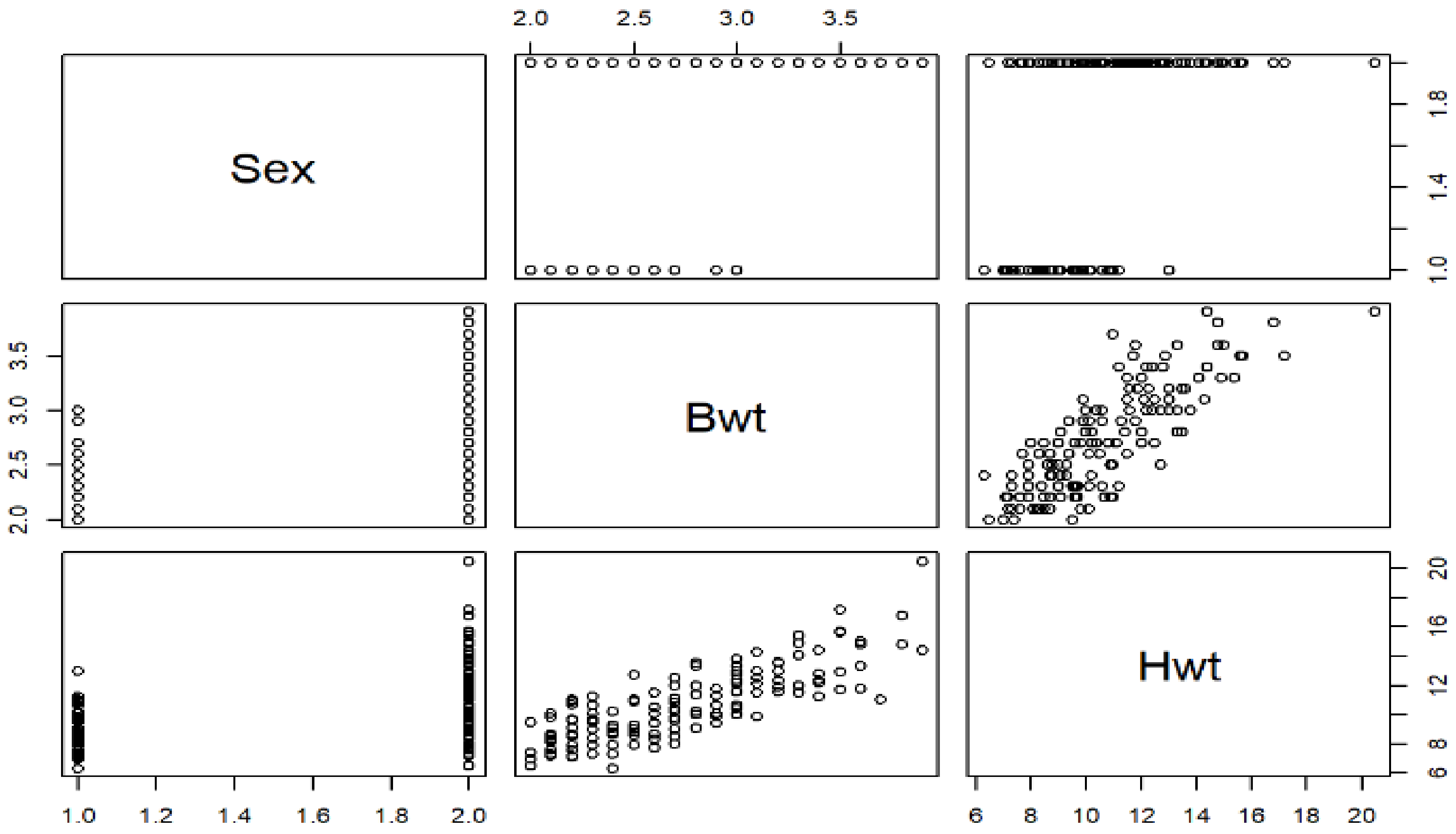
```
##  Sex         Bwt             Hwt       
##  F:47   Min.   :2.000   Min.   : 6.30  
##  M:97   1st Qu.:2.300   1st Qu.: 8.95  
##         Median :2.700   Median :10.10  
##         Mean   :2.724   Mean   :10.63  
##         3rd Qu.:3.025   3rd Qu.:12.12  
##         Max.   :3.900   Max.   :20.50  
```

cats 데이터 셋 불러오기
데이터 구조, 통계치 확인

실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
plot(cats)
```



데이터 선형성 확인을 위한 plot 그려보기

실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
model1 <- lm(cats$Hwt ~cats$Bwt, data = cats)
```

lm(종속변수 ~독립변수, data = 데이터명)

```
model1
```

```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt, data = cats)
##
## Coefficients:
## (Intercept)     cats$Bwt
##     -0.3567       4.0341
```

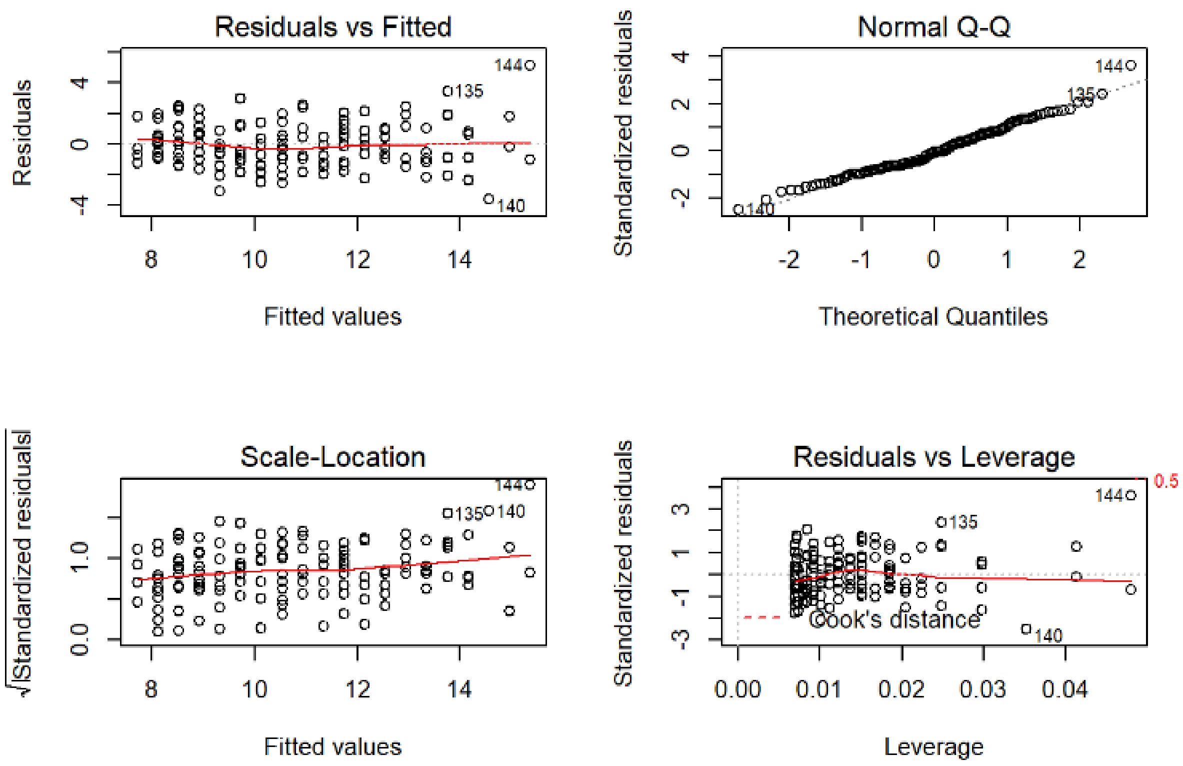
y = -0.3567 + 4.0341*x1(Bwt)

회귀식은 위와 같이 산출됐지만 과연 옳은 회귀식일까?

실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
par(mfrow = c(2,2))
plot(model1)
```



오차에 대한 가정의 검토

1. 등분산성 확인	2. 정규성 확인
• X축 : 회귀로 예측된 Y값 • Y축 : 잔차 ➢ 점들의 분포가 균일한 것이 이상적	➢ 점들의 분포가 기울기가 45인 직선을 따르는 모습을 보이는것이 이상적
3. 등분산성 확인(이상치)	**4. 이상치 확인**
• X축 : 회귀로 예측된 Y값 • Y축 : 표준화 잔차 ➢ 점들의 분포가 균일한 것이 이상적 ➢ 특정점이 0에서 멀리 떨어져있다면, 이상치일 가능성	• X축 : Leverage • Y축 : 표준화 잔차 ➢ 설명변수(특정점)가 얼마나 극단에 치우쳐 있는지를 확인

실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
summary(model1)
```

```
##
## Call:
## lm(formula = cats$Hwt ~ cats$Bwt, data = cats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5694 -0.9634 -0.0921  1.0426  5.1238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## cats$Bwt      4.0341     0.2503  16.119   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.452 on 142 degrees of freedom
## Multiple R-squared:  0.6466, Adjusted R-squared:  0.6441
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

명칭	설명
Residuals	잔차의 4분위수 범위
Coefficients	
(Intercept)	회귀식의 절편
Estimate	각 독립변수의 기울기 값
Std.error	표준오차
Pr(>|t|)	독립변수의 Estimate에 대한 P-value • 0.05 미만일 경우 통계적으로 유의함 • 0.05 이상일 경우 해당 독립변수를 채택하지 않음.

실습 - 단순선형회귀

1단계 : 회귀모형은 타당한가?

```
## F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Ho : 회귀모형은 타당하지 않다
H1 : 회귀모형은 타당하다

유의확률이 <2.2E-16 이므로
유의 수준 0.05에서 회귀모형은 통계적으로 타당하다.

즉, 독립변수가 영향을 준다.

2단계 : 독립변수들은 종속변수에게 영향을 주는가?
(조건 :1단계의 결론이 대립가설이어야 함

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## cats$Bwt      4.0341     0.2503  16.119   <2e-16 ***
## ---
```

Ho : 독립변수는 종속변수에게 영향을 주지 않는다
H1 : 독립변수는 종속변수에게 영향을 준다.

1. Bwt의 유의확률은 <2e-16 이므로
종속변수에 영향을 줌

Intercept의 유의확률 > 0.05 이지만,
절편값에 대한 유의성 검정은 귀무가설을
기각하지 못 해도 상관없다.

"절편 = 0" 이라는 귀무 가설의 경우 일반적으로 회귀분석의 관심 대상이 아니기 때문이다.

실습 - 단순선형회귀

3단계 : 독립변수는 종속변수에게 어떤 영향을 주는가?

```
""
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3567     0.6923  -0.515    0.607
## cats$Bwt      4.0341     0.2503  16.119   <2e-16 ***
## ---
```

Bwt 의 회귀계수는 4.0341 이므로
독립변수 기본단위가 1 증가하면,
종속변수는 약 4.0341 정도 시키는 영향을 준다.

y = -0.3567 + 4.0341 x1(Bwt)

4단계 : 회귀모형의 설명력(독립변수의 설명력)

```
## Multiple R-squared:  0.6466,
```

R-sqaured(결정계수) :
SSR/SST = 회귀모형의 설명력

-> 회귀모형의 설명력은 약 64.66% 정도 이다.

실습 - 단순선형회귀

Lorem Ipsum is simply dummy text of the printing and typesetting industry


```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: cats$Hwt
##            Df Sum Sq Mean Sq F value    Pr(>F)
## cats$Bwt    1 548.09  548.09  259.83 < 2.2e-16 ***
## Residuals 142 299.53    2.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

모델1의 ANOVA(분산분석) 테이블

변수의 F value
P-value 산출 과정 확인 가능

<단순회귀의 분산분석표>

요인	제곱합	자유도	평균제곱	F_0
회귀	SSR	1	$MSR=SSR$	$\frac{MSR}{MSE}$
잔차	SSE	$n-2$	$MSE=\frac{SSE}{n-2}$	
계	SST	$n-1$		

Sum sq : 제곱합을 나타냄

cats$Bwt ~ Sum Sq : SSR(회귀 변동)
Residuals ~ Sum Sq : SSE(오차 변동)
SSR + SSE = SST(오차의 총 제곱합)

실습 - 다중회귀분석

Lorem Ipsum is simply dummy text of the printing and typesetting industry

state.x77: 1977년 미국의 각 주에 대한 통계량	
Murder	살인사건 발생율
Population	인구
Illiteracy	문맹률
Income	수익
Frost	결빙일수

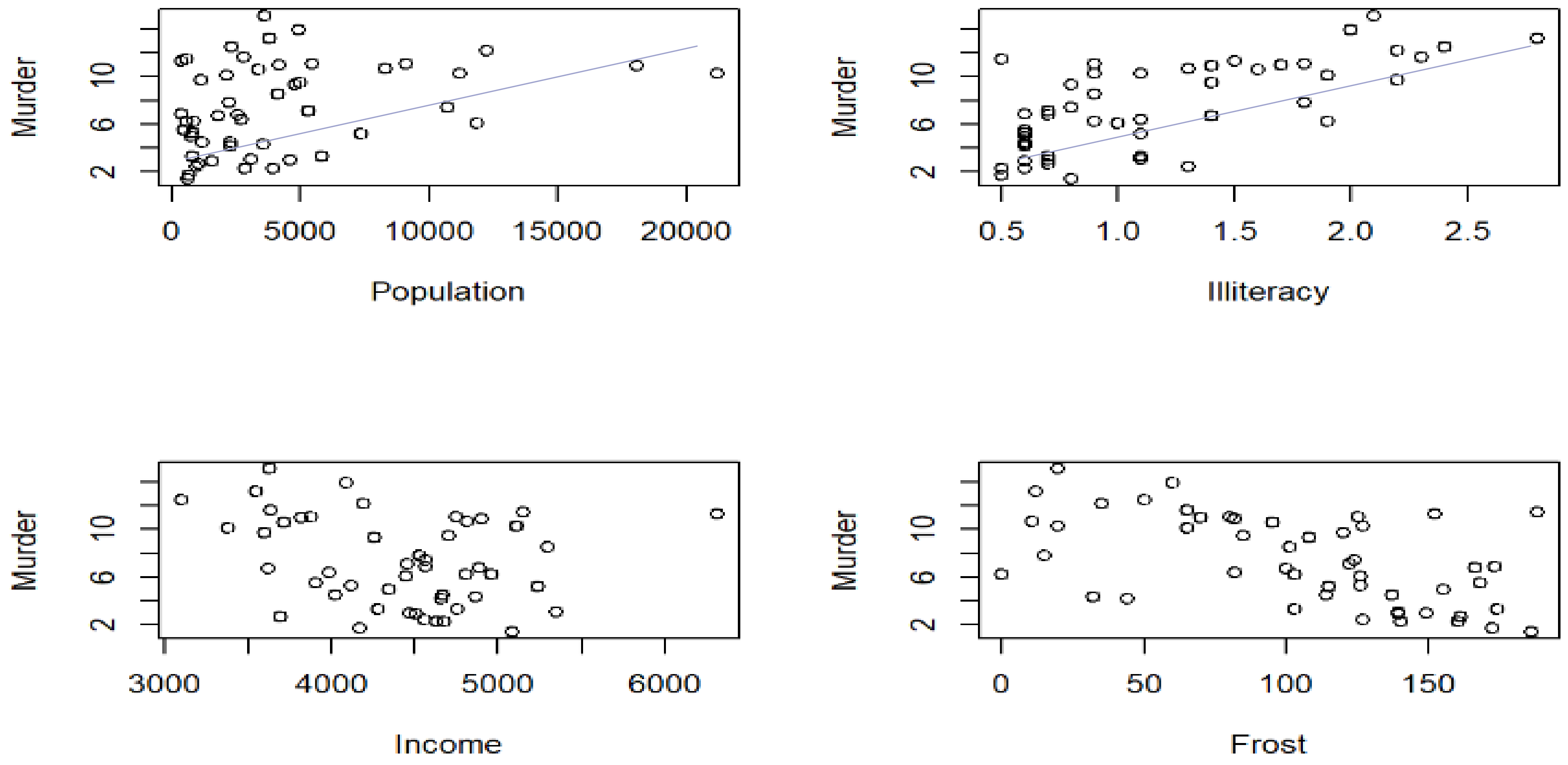
```
states <- as.data.frame(state.x77[,c("Murder", "Population", "Illiteracy", "Income","Frost")])
head(states)
```

```
##            Murder Population Illiteracy Income Frost
## Alabama      15.1       3615        2.1   3624    20
## Alaska       11.3        365        1.5   6315   152
## Arizona       7.8       2212        1.8   4530    15
## Arkansas     10.1       2110        1.9   3378    65
## California   10.3      21198        1.1   5114    20
## Colorado      6.8       2541        0.7   4884   166
```

실습 - 다중회귀분석

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
par(mfrow = c(2,2))
plot(Murder ~., data =states)
```



Y = Murder 일 경우

선형성 확인을 위한 plot 그리기

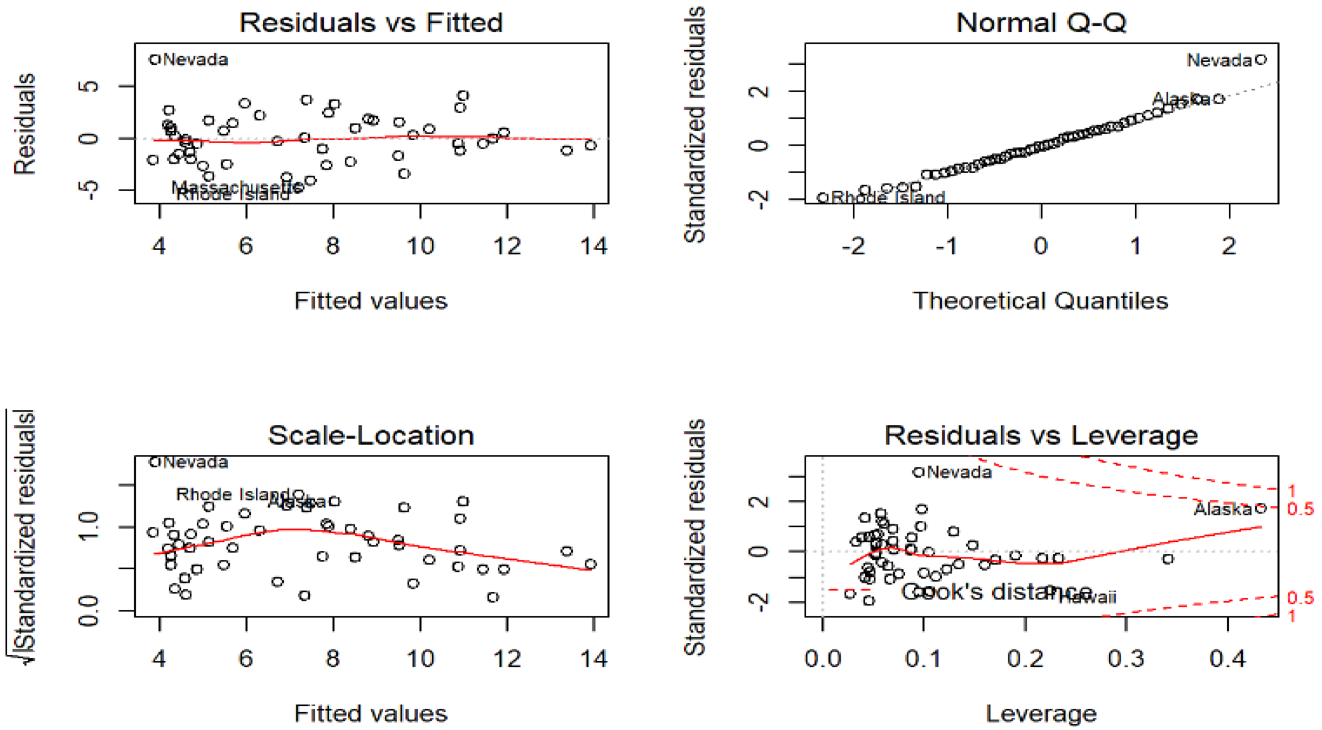
실습 - 다중회귀분석

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
model2 <- lm(Murder~Population+Illiteracy+Income+Frost, data = states)
model2 <- lm(Murder~., data = states) # 위에랑 같은 표현

lm(Y~x1 + x2 + x3 + x4 , data = data name) = lm(Y~., data= data name)

plot(model2)
```



오차에 대한 가정의 검토

1. 등분산성 검정	2. 정규성
점들이 고루 퍼져 있는 것으로 보임	처음과 끝은 완전한 정규성을 보인다 할 수 없음
3. 등분산성(이상치)검정	4. 이상치 확인
대부분의 점들은 고르게 퍼져있지만, 0으로부터 거리가 먼 이상치 확인.	Nevada, Alaska 등의 이상치가 확인됨

실습 - 다중회귀분석

Lorem Ipsum is simply dummy text of the printing and typesetting industry

```
summary(model2)
```

```
##
## Call:
## lm(formula = Murder ~ ., data = states)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7960 -1.6495 -0.0811  1.4815  7.6210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.235e+00  3.866e+00   0.319   0.7510
## Population  2.237e-04  9.052e-05   2.471   0.0173 *
## Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
## Income      6.442e-05  6.837e-04   0.094   0.9253
## Frost       5.813e-04  1.005e-02   0.058   0.9541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 45 degrees of freedom
## Multiple R-squared:  0.567,  Adjusted R-squared:  0.5285
## F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

실습 - 다중회귀분석

1단계 : 회귀모형은 타당한가?

```
## F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

Ho : 회귀모형은 타당하지 않다
H1 : 회귀모형은 타당하다

유의확률이 0.133e-08 이므로
유의 수준 0.05에서 회귀모형은 통계적으로 타당하다.

➢ 최소 한개 이상의 독립변수가 종속변수에게 영향을 끼친다.
(단순회귀 해석과 다른점)

2단계 : 독립변수들은 종속변수에게 영향을 주는가?
(조건 :1단계의 결론이 대립가설이어야 함

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.235e+00  3.866e+00   0.319   0.7510
## Population  2.237e-04  9.052e-05   2.471   0.0173 *
## Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
## Income      6.442e-05  6.837e-04   0.094   0.9253
## Frost       5.813e-04  1.005e-02   0.058   0.9541
```

Ho : 독립변수는 종속변수에게 영향을 주지 않는다
H1 : 독립변수는 종속변수에게 영향을 준다.

➢ Income , Frost 변수들은 유의확률 > 0.05 이므로 귀무가설을 기각하지 못함

실습 - 다중회귀분석

3단계 : 유의하지 않은 변수 제거 후 다중선형회귀식 만들기

```
model3 <- lm(Murder~Population + Illiteracy, data = states)
summary(model3)
```

```
##
## Call:
## lm(formula = Murder ~ Population + Illiteracy, data = states)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7652 -1.6561 -0.0898  1.4570  7.6758
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
## Population  2.242e-04  7.984e-05   2.808  0.00724 **
## Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.481 on 47 degrees of freedom
## Multiple R-squared:  0.5668, Adjusted R-squared:  0.5484
## F-statistic: 30.75 on 2 and 47 DF,  p-value: 2.893e-09
```

4단계 : 회귀모형 타당성 확인 및 독립변수들은 종속변수에게 영향성 확인


```
## F-statistic: 30.75 on 2 and 47 DF,  p-value: 2.893e-09
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
## Population  2.242e-04  7.984e-05   2.808  0.00724 **
## Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
```


1. 회귀모형의 유의확률(2.893e-09) < 0.05 이므로
   회귀모형은 통계적으로 유의하다

2. Population 유의확률(0.00724) < 0.05
   illiteracy 유의확률(8.83e-09) < 0.05 이므로
   두 변수는 종속변수에게 영향을 준다.

실습 - 다중선형회귀

5단계 : 독립변수들은 종속변수에게 어떤 영향을 주는가?

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
## Population  2.242e-04  7.984e-05   2.808  0.00724 **
## Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
## ...
```

Bwt 의 회귀계수는 4.0341 이므로
독립변수 기본단위가 1 증가하면,
종속변수는 약 4.0341 정도 시키는 영향을 준다.

y = 1.652e+00 + 2.242e-04 x1(Population) + 4.081e+00 x2 (illiteracy)

x1의 단위가 1 증가하면, + 2.242e-04 만큼 종속변수(살인율)에 영향을 줌.
x2의 단위가 1 증가하면, + 4.081e+00 만큼 종속변수(살인율)에 영향을 줌

6단계 : 회귀모형의 설명력(독립변수의 설명력)

```
## Multiple R-squared:  0.5668 Adjusted R-squared:  0.5484
```

다중회귀모형의 설명력을 볼때는
Adjusted R-sqaured(수정된 회귀계수) 값으로 확인

- 회귀변수의 개수가 많을수록 R-square의 값이 증가하므로 잘 못된 해석력을 보일 수 있기 때문이다.

회귀모형의 설명력

- 회귀모형의 설명력은 약 54.84% 정도 이다

THANK YOU

ㅎ,.ㅎ 7