WeDataLab

디랩컨설팅 디랩아카데미

단순 선형회귀 분석

단순 선형회귀분석이란?

정의 종속 변수 Y와 하나의 독립 변수 X와의 선형 상관 관계를 모델링하는 회귀분석기법이다.

목적 단순 선형회귀분석을 이해하고, R을 활용하여 **실습**해보자

Goal!

0. 회귀(Regression)의 뜻

1. 단순 선형회귀분석은 어떻게 **활용**될 수 있는가?

2. 단순 선형회귀분석의 **개념**을 알아보자

3. 단순 선형회귀분석 : **모형의 적합도 평가 <- 잔차검정**

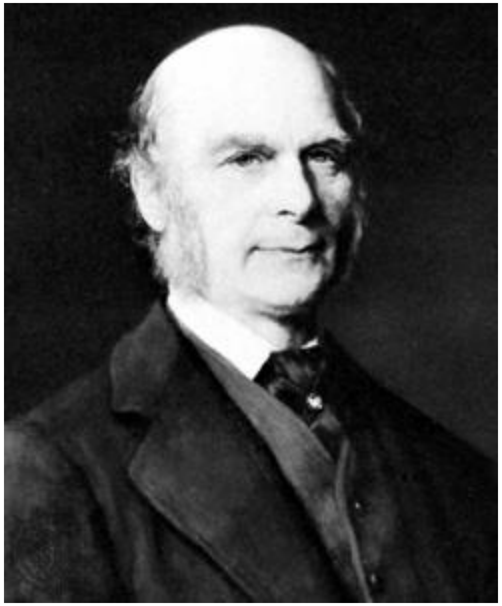
4. 예제 R 코드를 통해 단순 선형회귀분석을 **실습**해보자

1. 회귀의 뜻과 용어
2. 회귀계수 찾는 법
3. 회귀식 해석하기
4. 모형의 적합도 평가
5. 실습
6. 잔차분석
7. 회귀분석 다시하기

회귀(Regression) 의 뜻

프랜시스 골턴 (Francis Galton)

칼 피어슨 (Karl Pearson)



(1822년 ~ 1911년)

설명 변수(Explanatory Variable)

- 종속 변수에 선행하며, 종속 변수에 영향을 줄 것으로 기대되거나 종속 변수의 변화를 예측(predict)할 수 있다고 여겨지는 변수

- 원인으로 간주되는 변수

- 독립 변수 (independent variable), 예측 변수 (predictor variable), 입력 변수 (input variable), 특징 (feature) 등 여러 이름으로 부른다.

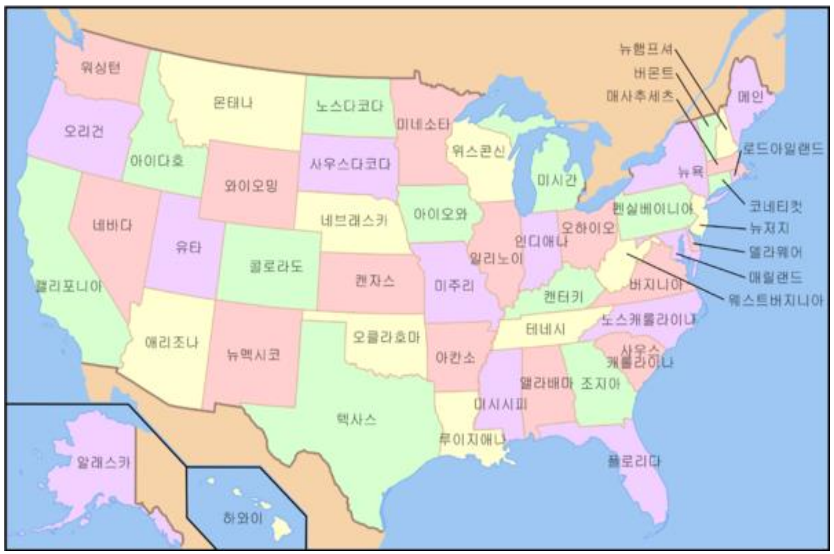
회귀분석이란

반응 변수(Response Variable)

- 독립 변수의 변화에 의해 영향을 받을 것으로 기대되는 변수
- 결과로 간주되는 변수
- 종속변수(**Dependent Variable**)라고도 한다.

Goal 1. 단순 선형회귀분석은 어떻게 활용될 수 있는가?

1977년 미국의 각 주의 대한 통계량 사례



< 미국의 50개 주 >

종속 변수

변수	변수 설명
Murder	인구 10만명 당 살인사건 수
Population	인구수
Income	수입
illiteracy	문맹률
Life Exp	기대 수명
HS Grad	고등학교 졸업률
Frost	섭씨 0도 이하로 내려간 날의 수
Area	주의 면적

설명 변수 후보들(이 중 하나 선택)

변수 선택법

- 관심 있는 데이터에 대한 선행연구
- 이론적인 배경
- 분석가의 직관

위 사항들을 종합적으로 고려하여 종속 변수 y와 설명 변수 x를 결정

이게 뒤바뀌면 …

회귀분석이란

회귀분석의 기본가정

두 변수의 관계는 선형이다. (선형성)

오차항의 확률 분포는 정규분포를 이루고 있다. (정규성)

오차항은 모든 독립변수 값에 대하여 동일한 분산을 갖는다. (등분산성)

오차항의 평균(기대값)은 0이다. (*Zero-conditional Mean)*

오차항들끼리는 독립이다. 어떤 패턴을 나타내면 안 된다. (독립성)
독립변수 상호간에는 상관관계가 없어야 한다. <- Multicollinearity

https://brunch.co.kr/@gimmesilver/17 참고

상관분석와 회귀분석

상관은 인과관계와 무관하다.

회귀분석은 인과관계를 가정한다.

https://brunch.co.kr/@gimmesilver/17 참고

회귀분석의 목표

$$\boxed{\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x}$$

회귀모형

$$y = \beta_0 + \beta_1 x + \varepsilon$$

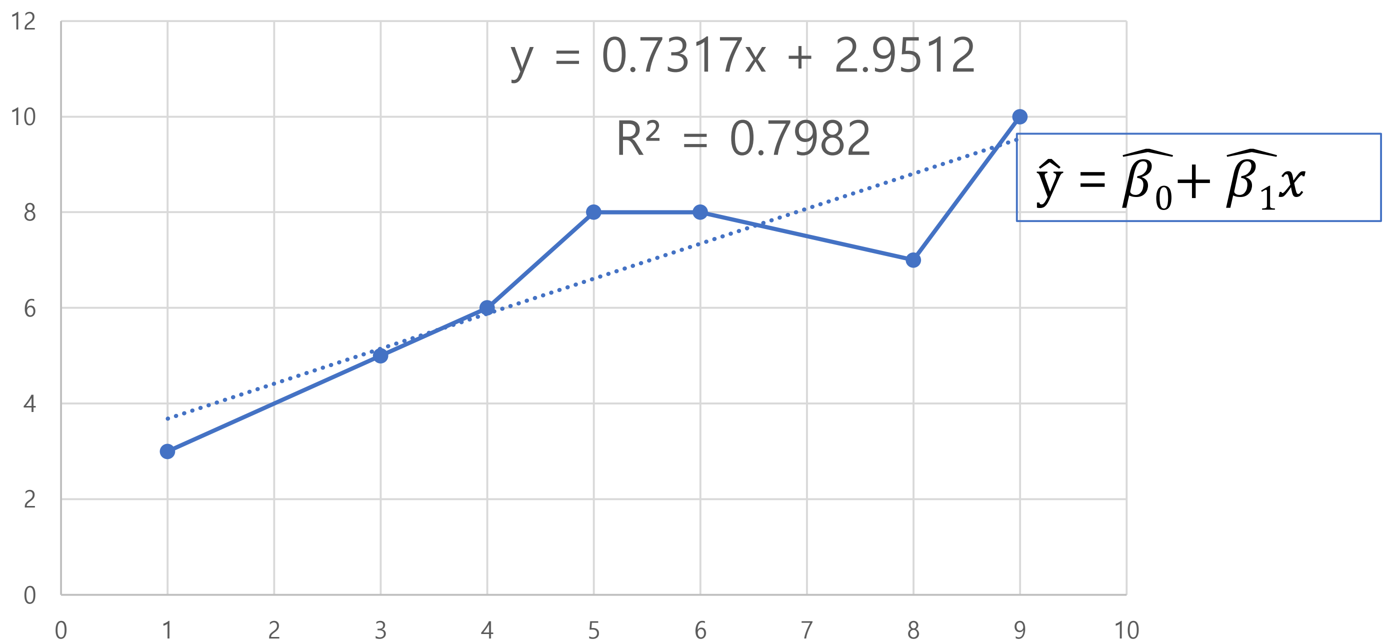
독립 변수와 종속 변수의 관계를 찾는 것

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$

종속변수 y 와 설명변수 x사이의 관계를 선형으로 가정하고

이를 가장 잘 설명하는 회귀 계수(coefficients)를 추정하는 것

회귀분석의 목표



- 회귀계수 β_1 은 설명변수 x가 한 단위(1) 증가할 때 종속변수가 얼마나 변화하는지 상관관계(correlation)를 보여주는 지표

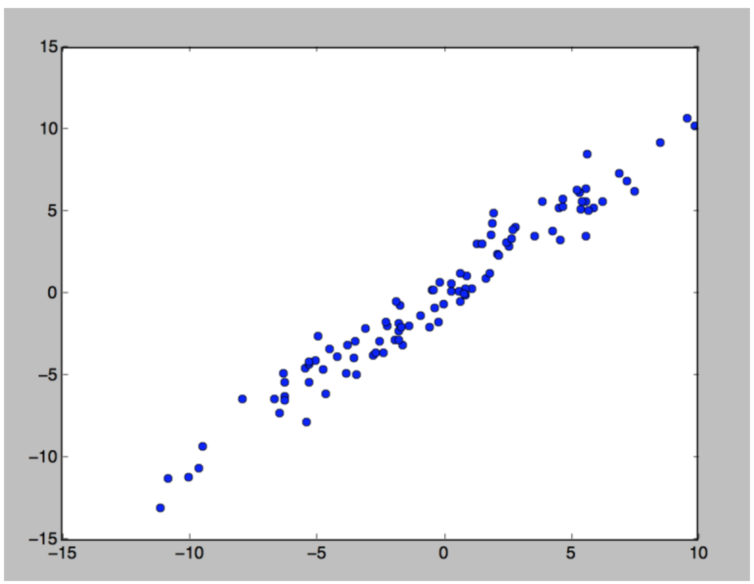
Goal 2. 단순 선형회귀분석의 개념을 알아보자

선형 회귀분석(Linear Regression)

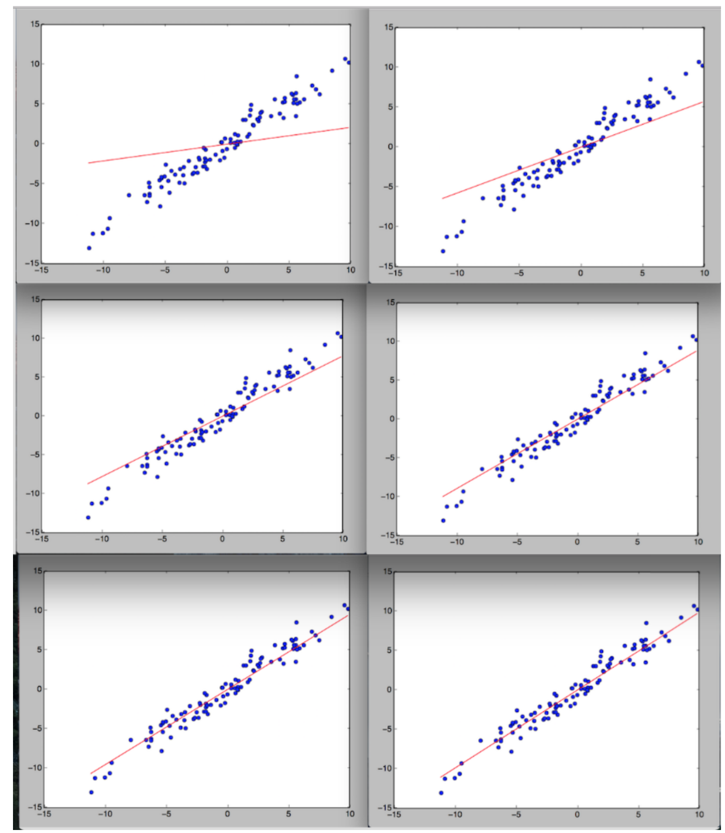
- 반응변수와 설명변수 사이의 관계를 선형으로 표현

단순선형회귀 예시

$$y = x + \varepsilon$$



점들을 가장 잘 적합하는
$\hat{\beta}$(직선)을 찾는 것



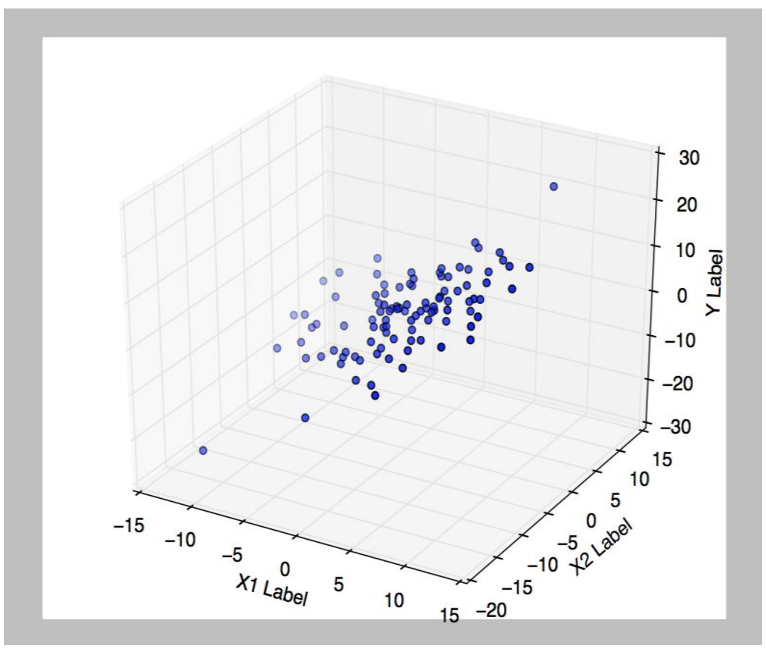
Goal 2. 단순 선형회귀분석의 개념을 알아보자

선형 회귀분석(Linear Regression)

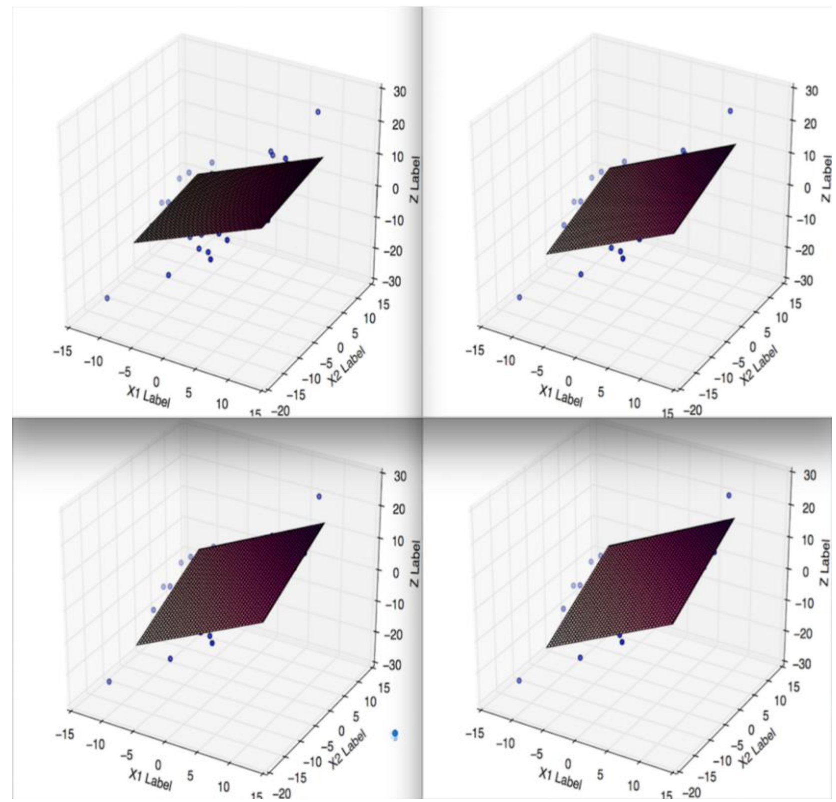
- 반응변수와 설명변수 사이의 관계를 선형으로 표현

다중선형회귀 예시

$$y = x_1 + x_2 + \varepsilon$$



점들을 가장 잘 적합하는
여러 개의 $\hat{\beta}$(평면)을 찾는 것



상관계수와 회귀계수의 관계

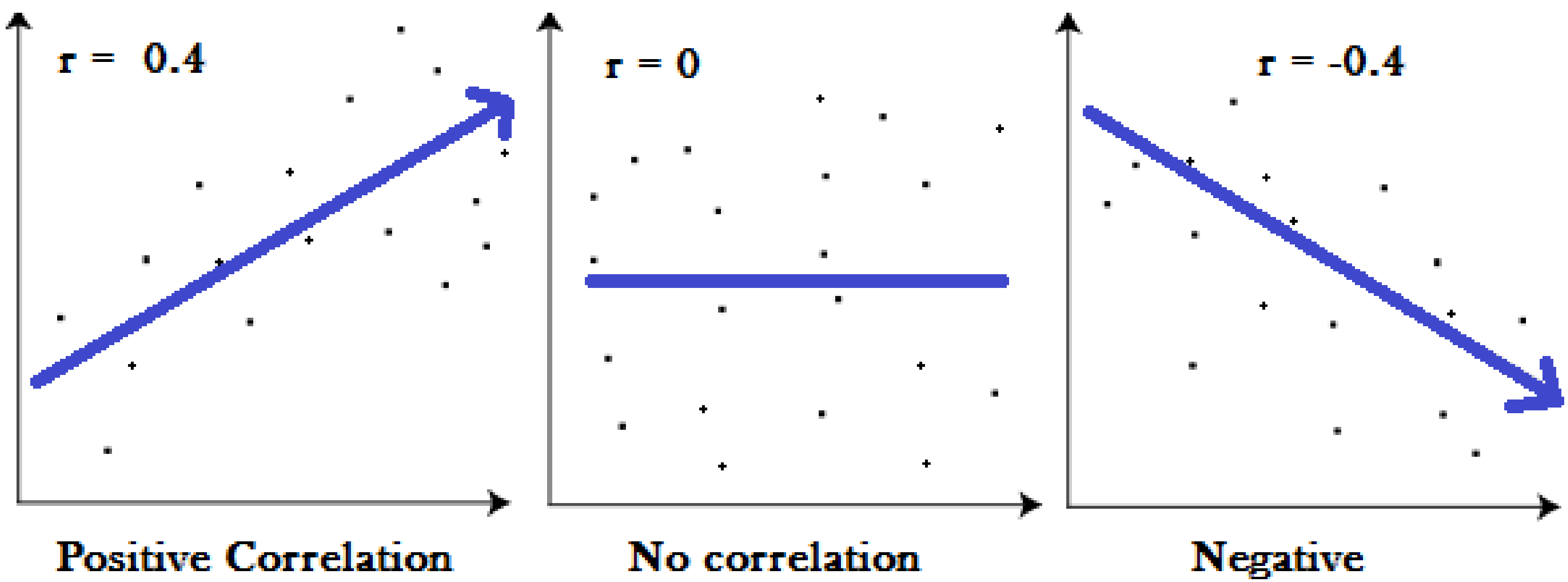
<u>표본상관계수의 성질</u>

(1) $-1 \leqq r \leqq 1$

(2) $|r|$의 값이 1에 가까울수록 강한 상관관계

$|r|$의 값이 0에 가까울수록 약한 상관관계

<u>여러 가지 경우의 산점도와 표본 상관계수</u>



source : http://www.statisticshowto.com/wp-content/uploads/2012/10/pearson-2-small.png, 서울대학교 통계학실험 강의자료

1. 회귀의 뜻과 용어
2. 회귀계수 찾는 법
3. 회귀식 해석하기
4. 모형의 적합도 평가
5. 실습
6. 잔차분석
7. 회귀분석 다시하기

회귀 계수의 추정

- 회귀분석의 목적: 종속변수 Y와 설명변수 X 사이의 관계를 선형으로 가정하고 이를 가장 잘 설명하는 회귀 계수(coefficients)를 추정하는 것

Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

- 회귀분석의 목적: 종속변수 Y와 설명변수 X 사이의 관계를 선형으로 가정하고 이를 가장 잘 설명하는 회귀 계수(coefficients)를 추정하는 것

-> 여기서 '잘 설명한다'는 의미는? 어떤 기준에서 '잘 설명'하는 것인가?

Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

- 회귀분석의 목적: 종속변수 Y와 설명변수 X 사이의 관계를 선형으로 가정하고 이를 가장 잘 설명하는 회귀 계수(coefficients)를 추정하는 것

  -> 여기서 '잘 설명한다'는 의미는? 어떤 기준에서 '잘 설명'하는 것인가?

  -> '적합된 직선'과 '실제 데이터' 사이에 'y축 거리(의 제곱)'의 합을 최소화

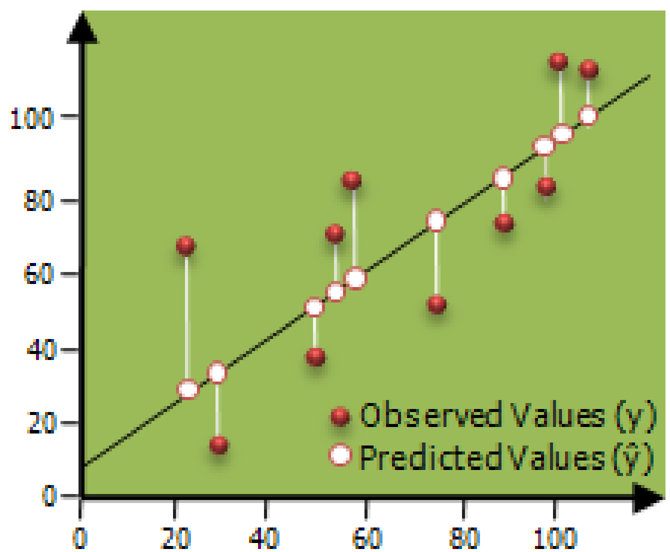
회귀 계수의 추정

- 회귀분석의 목적: 종속변수 Y와 설명변수 X 사이의 관계를 선형으로 가정하고 이를 가장 잘 설명하는 회귀 계수(coefficients)를 추정하는 것

-> 여기서 '잘 설명한다'는 의미는? 어떤 기준에서 '잘 설명'하는 것인가?

-> '적합된 직선'과 '실제 데이터' 사이에 'y축 거리(의 제곱)'의 합을 최소화

-> 최소제곱법(OLS; Ordinary Least Squares)



회귀 계수의 추정

- 최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

회귀 계수의 추정

최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

->ε은 오차항(error term)

(불확실성을 표현하는) 관측 불가능한 추상적인 개념

회귀 계수의 추정

최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

->ε은 오차항(error term), 관측 불가능한 추상적인 개념

예측된 y 값: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

회귀 계수의 추정

최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

->ε은 오차항(error term), 관측 불가능한 추상적인 개념

예측된 y 값: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

->ε을 관측할 수 없으므로 y와의 차이 존재

회귀 계수의 추정

최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

예측된 y 값: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

목적 : $\hat{y}$과 y의 **차이를 최소화**(정확히는 y축 거리의 제곱의 합)

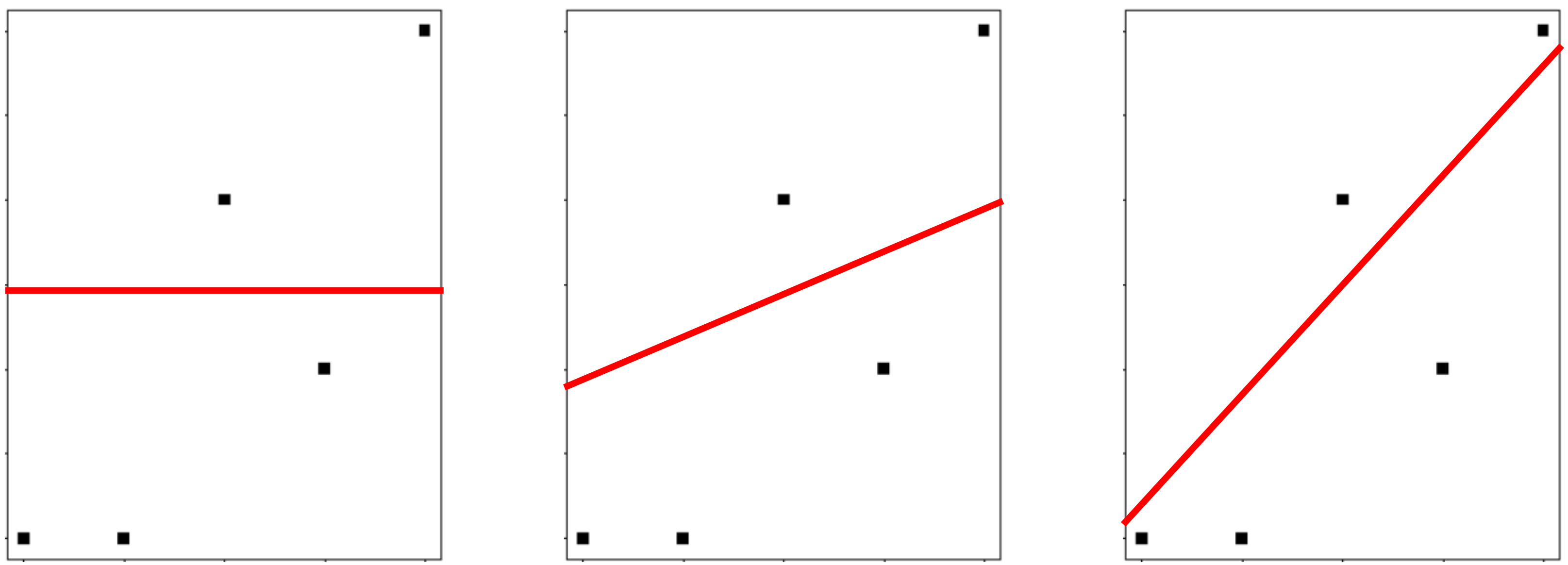
회귀 계수의 추정

최소제곱법(OLS; Ordinary Least Squares)

실제 y 값: $y = \beta_0 + \beta_1 x + \varepsilon$

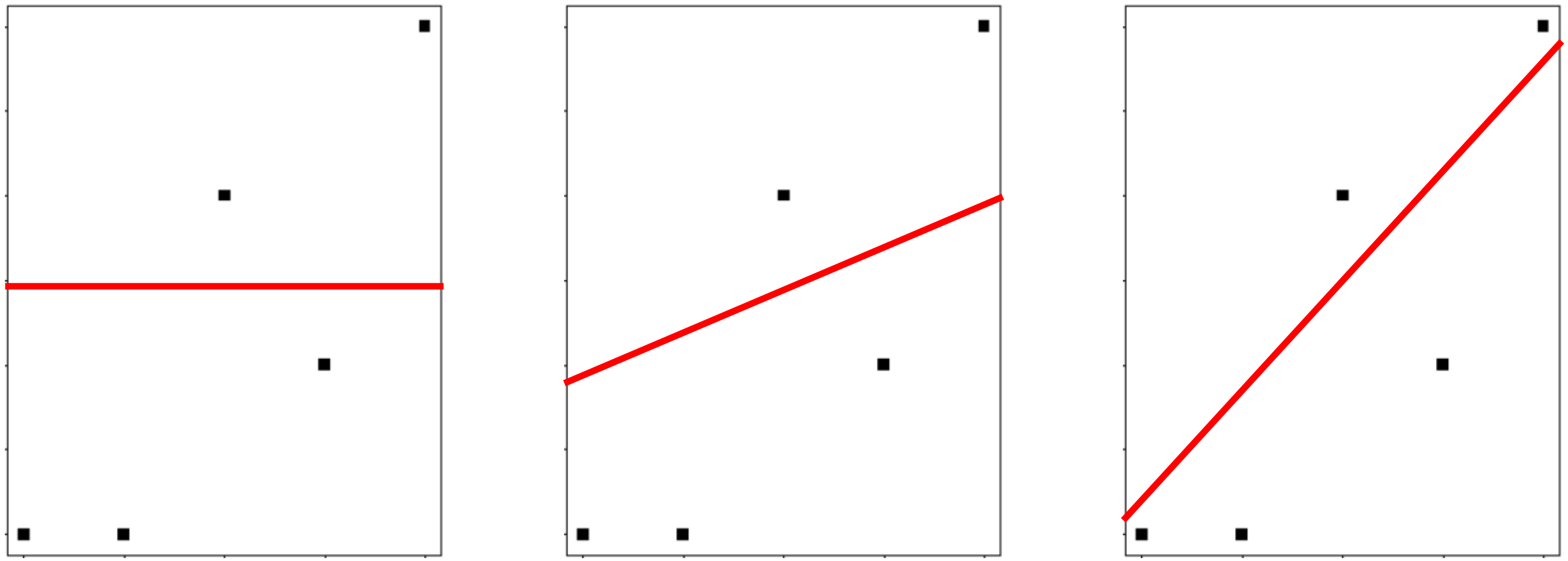
예측된 y 값: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

목적 : $\hat{y}$과 y의 차이를 최소화(정확히는 y축 거리의 제곱의 합)



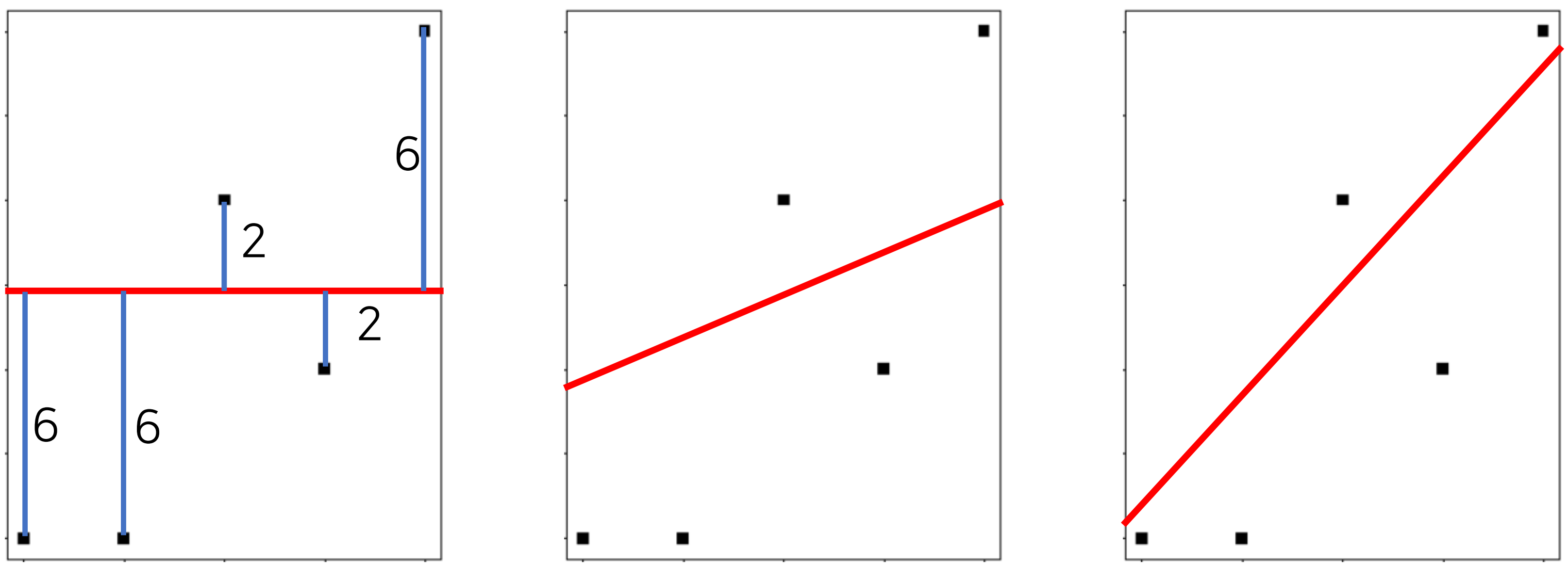
회귀 계수의 추정

최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화



회귀 계수의 추정

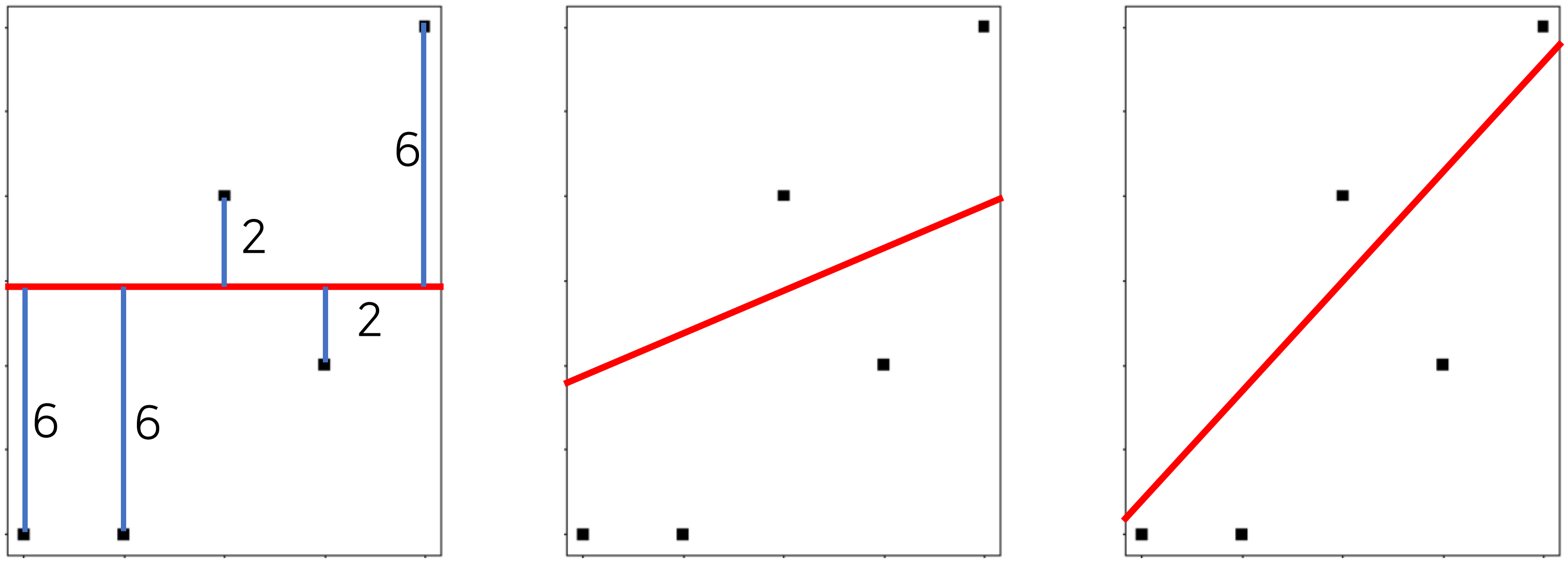
최소제곱법: $(\hat{y}_1-y_1)^2+(\hat{y}_2-y_2)^2+(\hat{y}_3-y_3)^2+(\hat{y}_4-y_4)^2+(\hat{y}_5-y_5)^2$ 을최소화



회귀 계수의 추정

최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

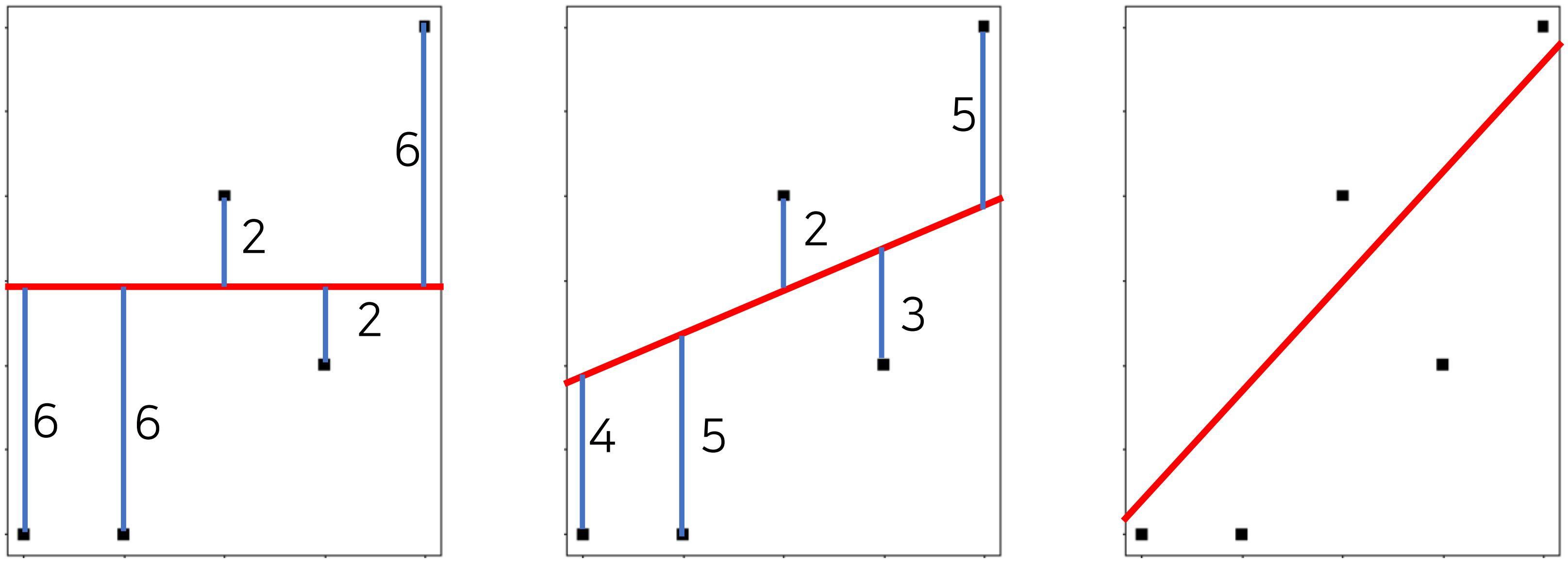
그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$



회귀 계수의 추정

- 최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$

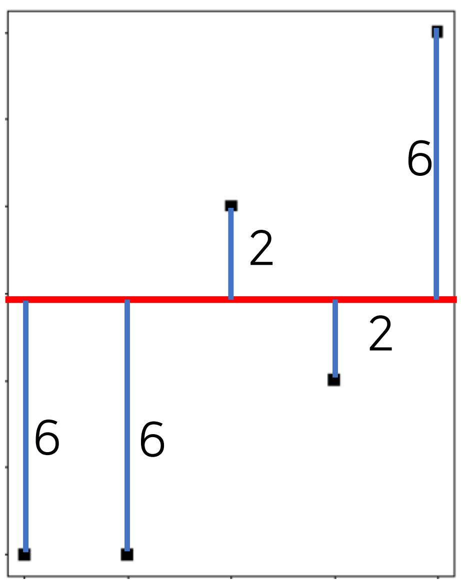


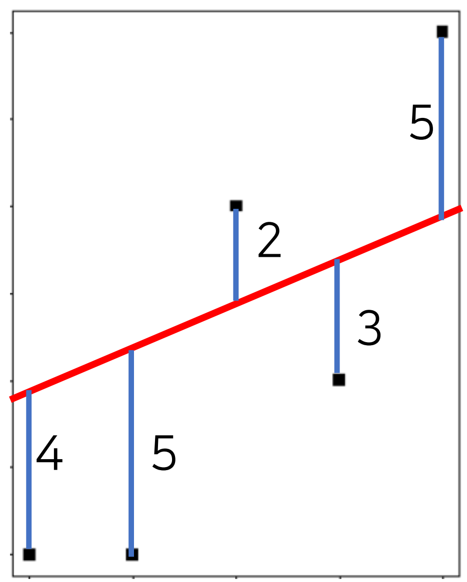
회귀 계수의 추정

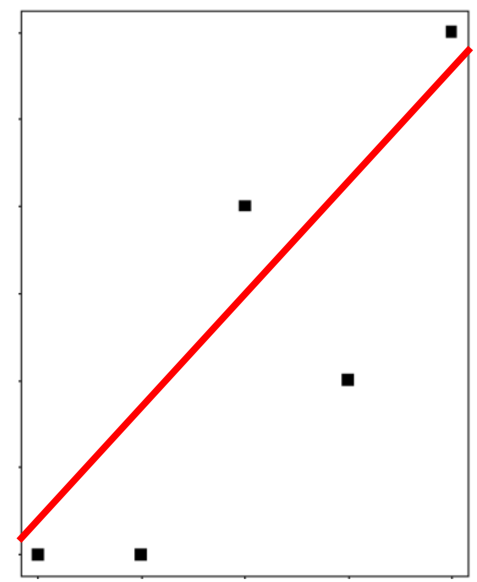
- 최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$

그림 2: $4^2 + 5^2 + 2^2 + 3^2 + 5^2 = 79$







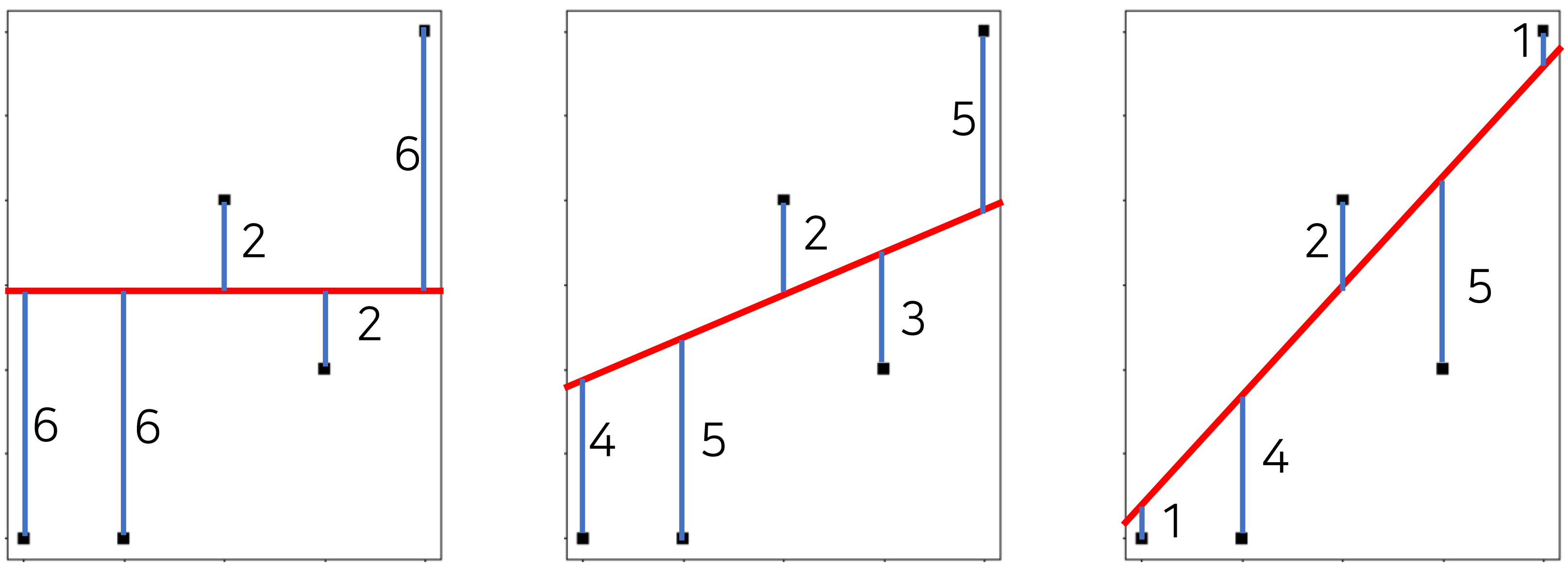
Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

- 최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$

그림 2: $4^2 + 5^2 + 2^2 + 3^2 + 5^2 = 79$



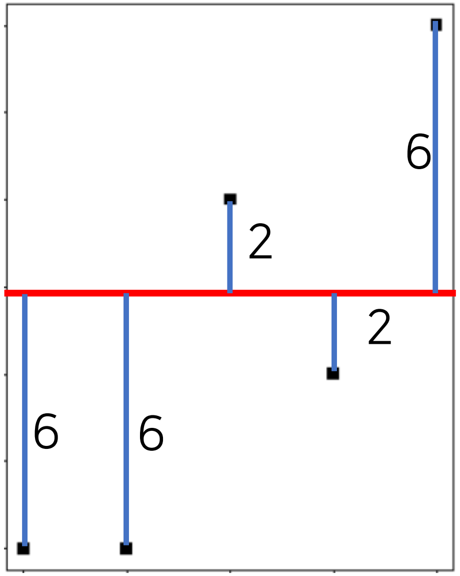
회귀 계수의 추정

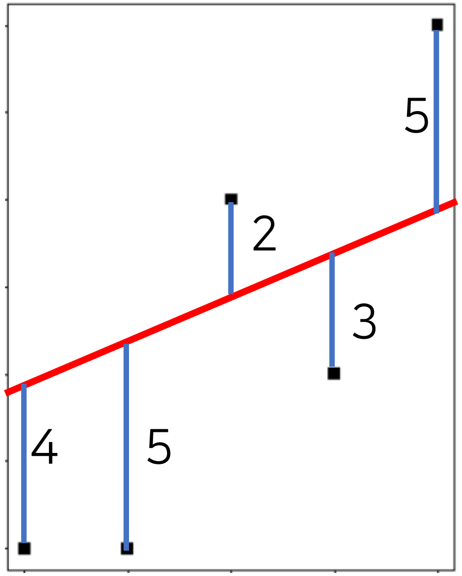
- 최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

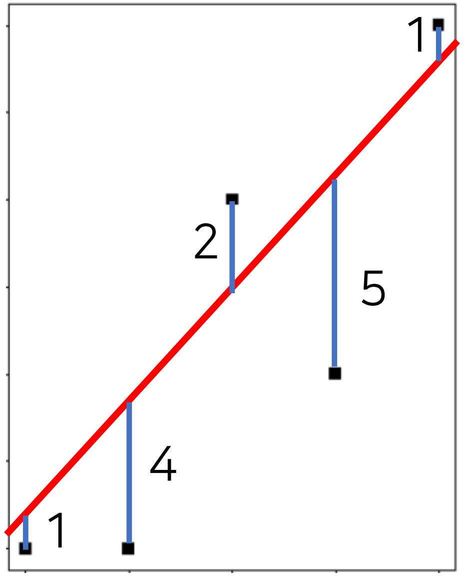
그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$

그림 2: $4^2 + 5^2 + 2^2 + 3^2 + 5^2 = 79$

그림 3: $1^2 + 4^2 + 2^2 + 5^2 + 1^2 = 47$







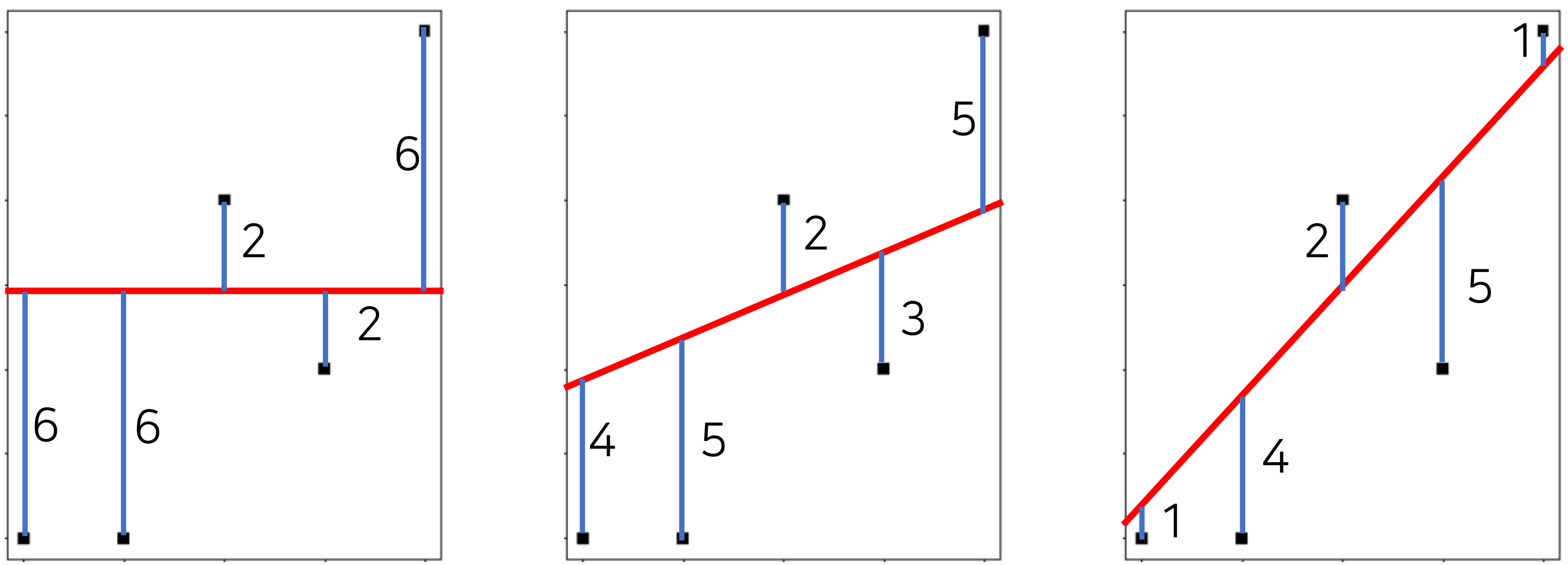
회귀 계수의 추정

최소제곱법: $(\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 + (\hat{y}_3 - y_3)^2 + (\hat{y}_4 - y_4)^2 + (\hat{y}_5 - y_5)^2$ 을최소화

그림 1: $6^2 + 6^2 + 2^2 + 2^2 + 6^2 = 116$

그림 2: $4^2 + 5^2 + 2^2 + 3^2 + 5^2 = 79$

그림 3: $1^2 + 4^2 + 2^2 + 5^2 + 1^2 = 47$ ->세 직선 중 그림 3의 직선을 선택



Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

요약: 최소제곱법(OLS; Ordinary Least Squares)

Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

요약: 최소제곱법(OLS; Ordinary Least Squares)

- 목적

\- 추정된 회귀식에 의해 결정된 값과 실제 종속변수 값의 차이를 최소화

회귀 계수의 추정

요약: 최소제곱법(OLS; Ordinary Least Squares)

- 목적

- 추정된 회귀식에 의해 결정된 값과 실제 종속변수 값의 차이를 최소화

- 방법 : 차이의 '제곱'을 최소화하도록 회귀계수 β_0 , β_1 를 추정

회귀 계수의 추정

요약: 최소제곱법(OLS; Ordinary Least Squares)

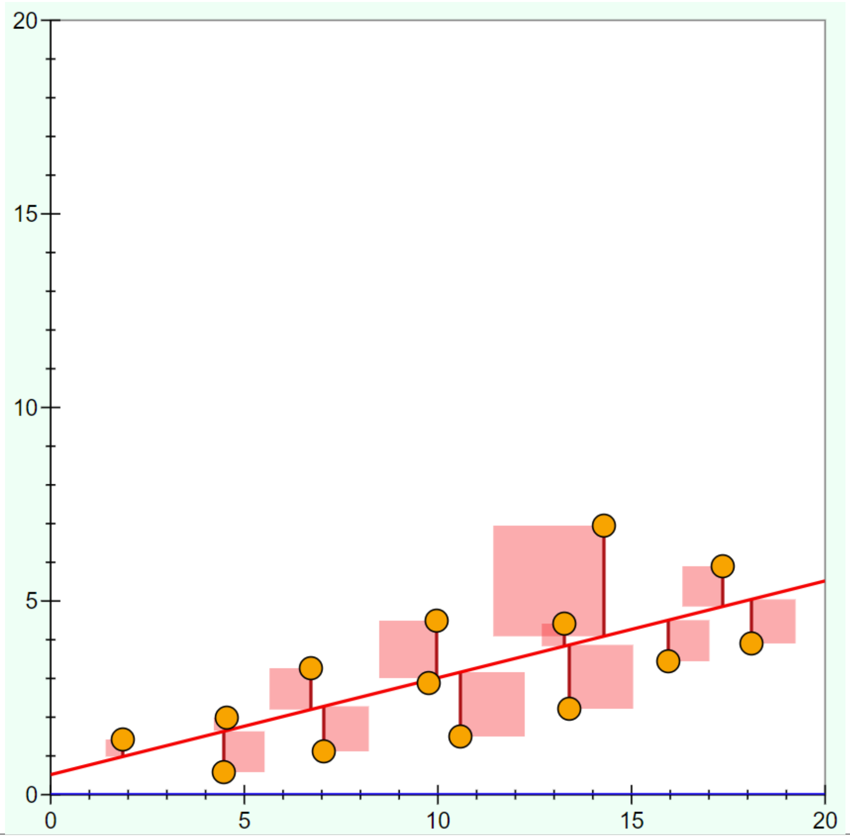
- 목적
- 추정된 회귀식에 의해 결정된 값과 실제 종속변수 값의 차이를 최소화

- 방법 : 차이의 '제곱'을 최소화하도록 회귀계수 β_0, β_1를 추정
- 보통 복잡한 함수를 사용하면 계수를 추정하는 것이 쉽지 않다.
- **'선형'** 회귀분석의 경우는 회귀계수가 closed form으로 한 번에 나옴

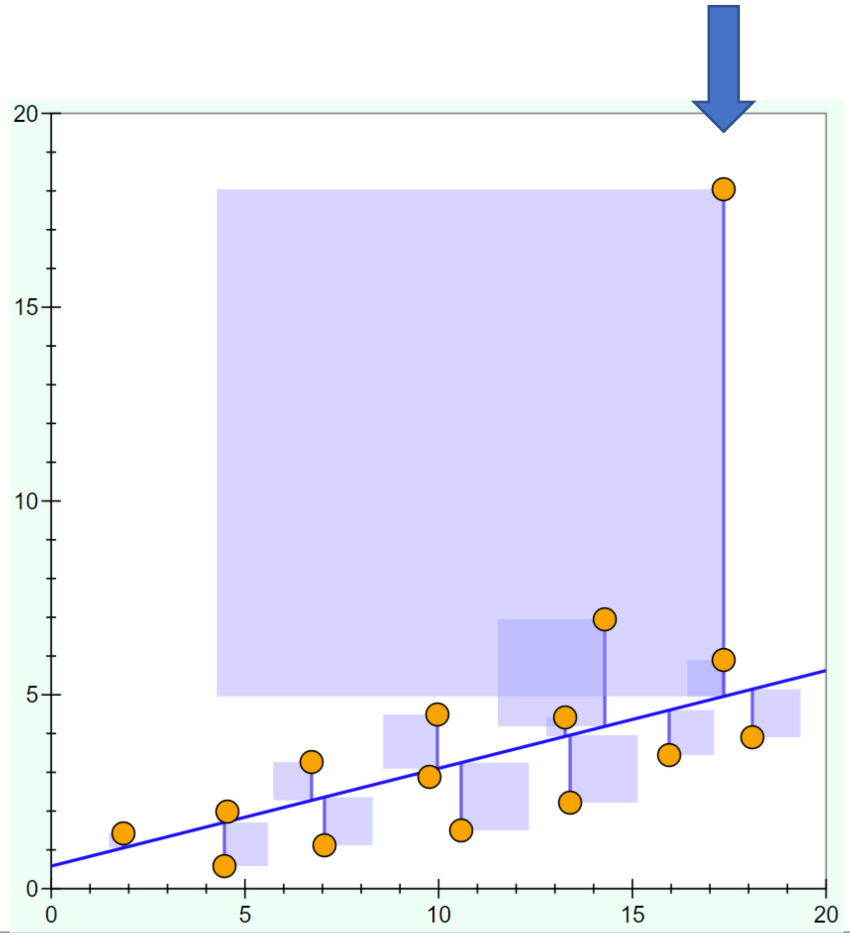
$$b_1 = \frac{s_{xy}}{s_x^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

잔차제곱합을 시각적으로 표현

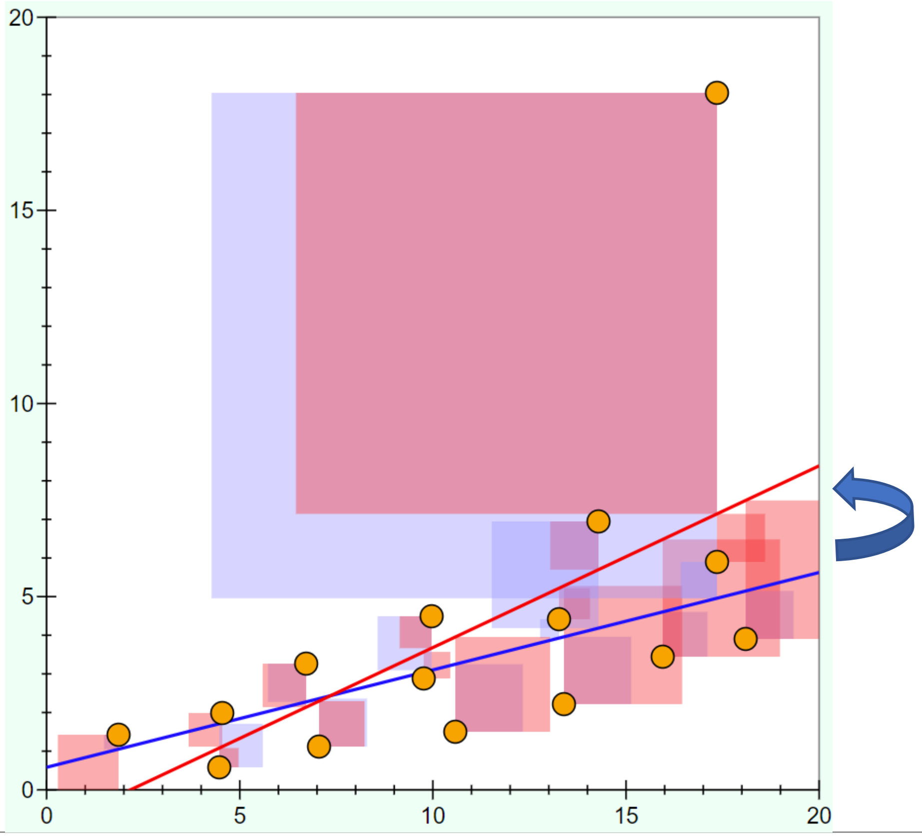
$$Cost(\hat{\beta}) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$



제곱오차를 사용하면 오차가 커질수록 비용이 엄청나게 커진다.



이상치(outlier)가 존재하면, 비용(잔차제곱합)을 줄이기 위해
직선이 이상치쪽으로 이동하는 경향



Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

(참고) 최소제곱법(OLS; Ordinary Least Squares)

지금까지 내용을 수식으로 표현하면 아래와 같다.

데이터 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ 이 주어진 경우:

아래의 최적화(최소화) 문제를 풀어 β_0, β_1을 구한다.

$$\min \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \quad (\beta_0, \beta_1 \in \mathbb{R})$$

Goal 2. 단순 선형회귀분석의 개념을 알아보자

회귀 계수의 추정

(참고) 최소제곱법(OLS; Ordinary Least Squares)

행렬을 이용해서 앞의 과정을 다시 쓰면 아래와 같다.

$$\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \Leftrightarrow \min\ (\mathrm{y} - \mathrm{X}\beta)^{\mathrm{T}}(\mathrm{y} - \mathrm{X}\beta)$$

여기서 $\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

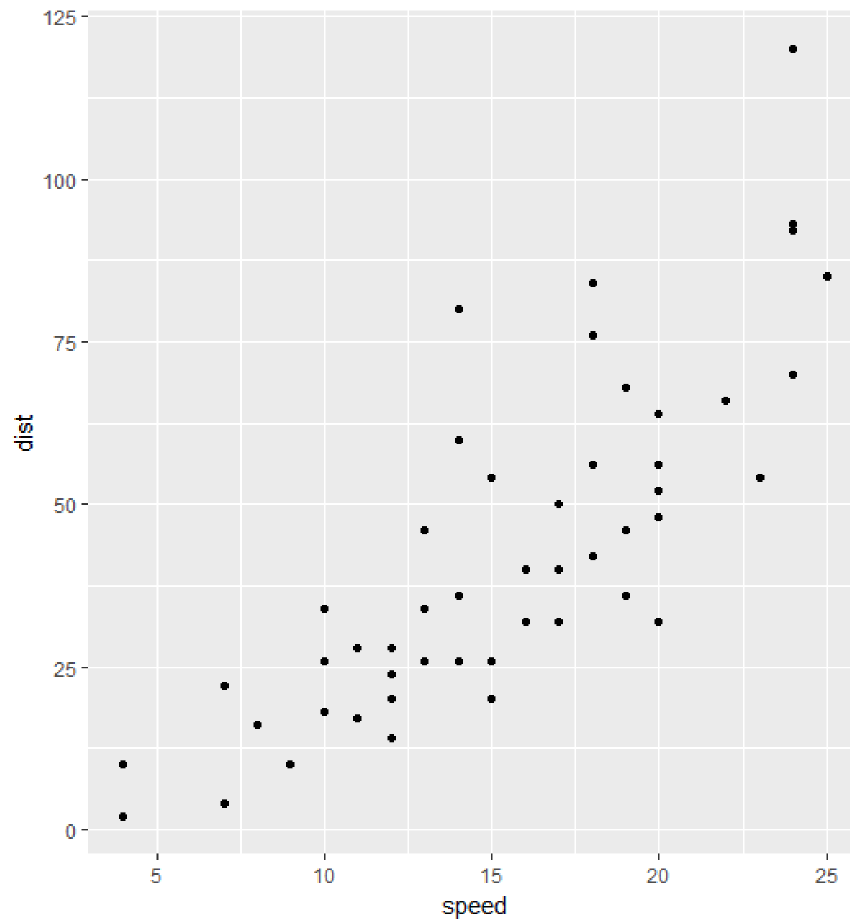
위의 최소화 문제를 풀면 $\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{y}$

1. 회귀의 뜻과 용어
2. 회귀계수 찾는 법
3. 회귀식 해석하기
4. 모형의 적합도 평가
5. 실습
6. 잔차분석
7. 회귀분석 다시하기

Goal 2. 단순 선형회귀분석의 개념을 알아보자

예시: “cars” 데이터(내장 데이터) : 브레이크가 작동되는 순간의 자동차의 주행 속도(Speed)에 따른 자동차 제동 거리(StopDist)를 조사한 자료

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26
21	14	36
22	14	60



Goal 2. 단순 선형회귀분석의 개념을 알아보자

예시: “cars” 데이터(내장 데이터) : 브레이크가 작동되는 순간의 자동차의 주행 속도(Speed)에 따른 자동차 제동 거리(StopDist)를 조사한 자료

```
> str(cars)
'data.frame':   50 obs. of  2 variables:
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

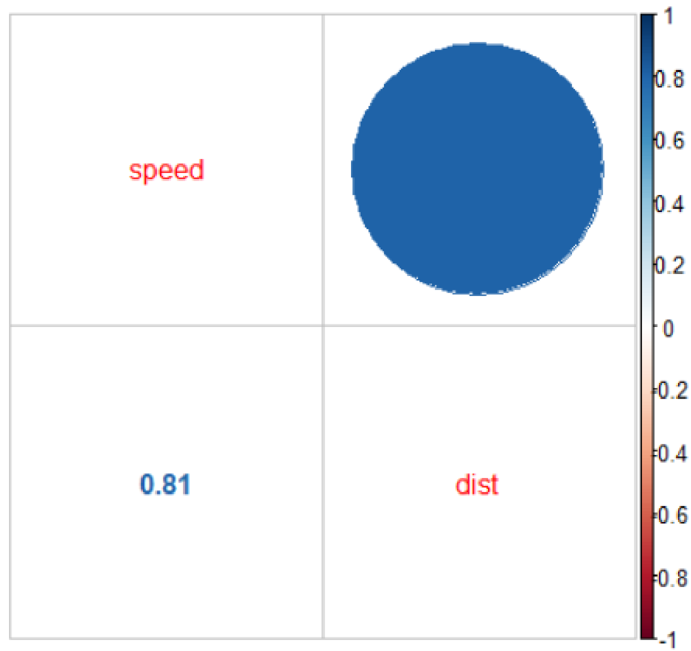
- speed: 자동차의 주행 속도
- dist: 자동차의 제동 거리
- 총 50개의 관측값

Goal 2. 단순 선형회귀분석의 개념을 알아보자

예시: “cars” 데이터(내장 데이터)

- 상관관계 확인: 회귀분석은 기본적으로 상관관계에 대한 통계적 분석
- 인과관계를 주장하려면 추가적으로 통계 외적인 근거가 필요

(예: 해당 분야의 이론적 근거, 실험의 경우 변인통제, …)



예시: “cars” 데이터(내장 데이터)

반응변수를 ‘dist’로, ‘speed’를 설명변수로 모형을 적합시킨 결과

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.5791	6.7584	-2.601	0.0123
speed	3.9324	0.4155	9.464	1.49e-12

예시: “cars” 데이터(내장 데이터)

Pr(>|t|): 유의확률(p-value)을 나타내며 확률이므로 0과 1 사이의 값

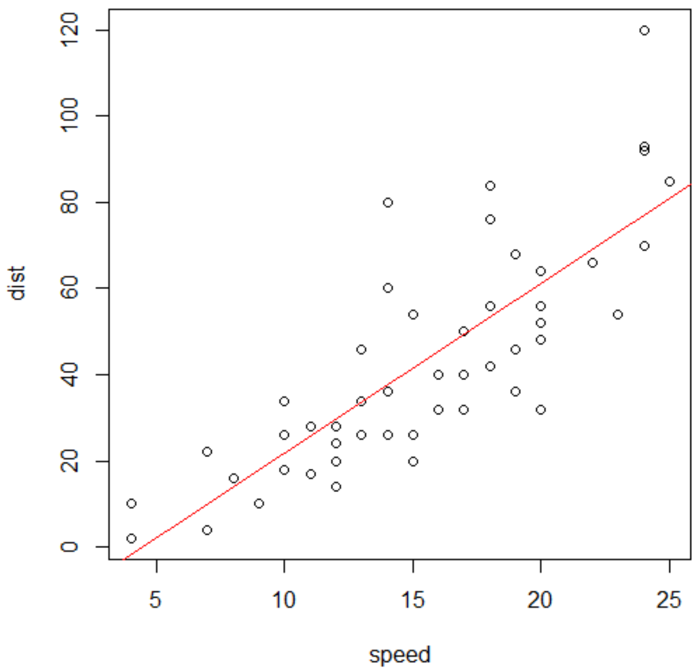
- 0에 가까울수록 중요한(통계적으로 유의미한) 변수임을 의미한다.
- speed 변수는 값이 매우 작으므로 유의하다.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.5791	6.7584	-2.601	0.0123
speed	3.9324	0.4155	9.464	1.49e-12

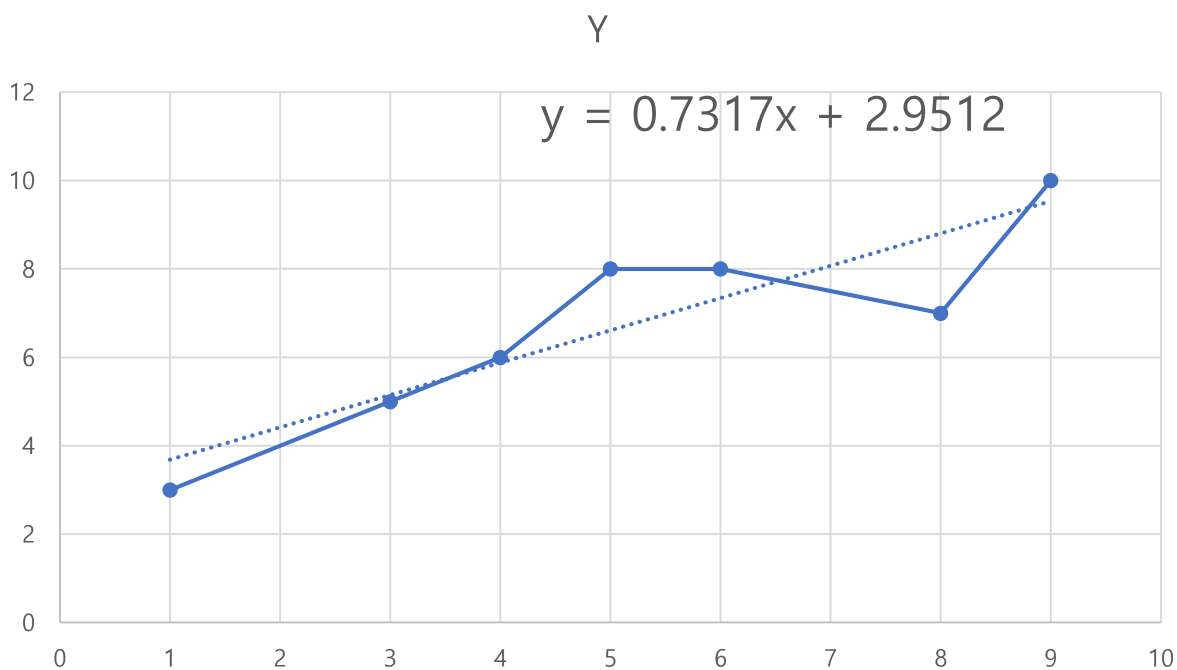
예시: “cars” 데이터(내장 데이터)

Estimate: 추정된 회귀계수값

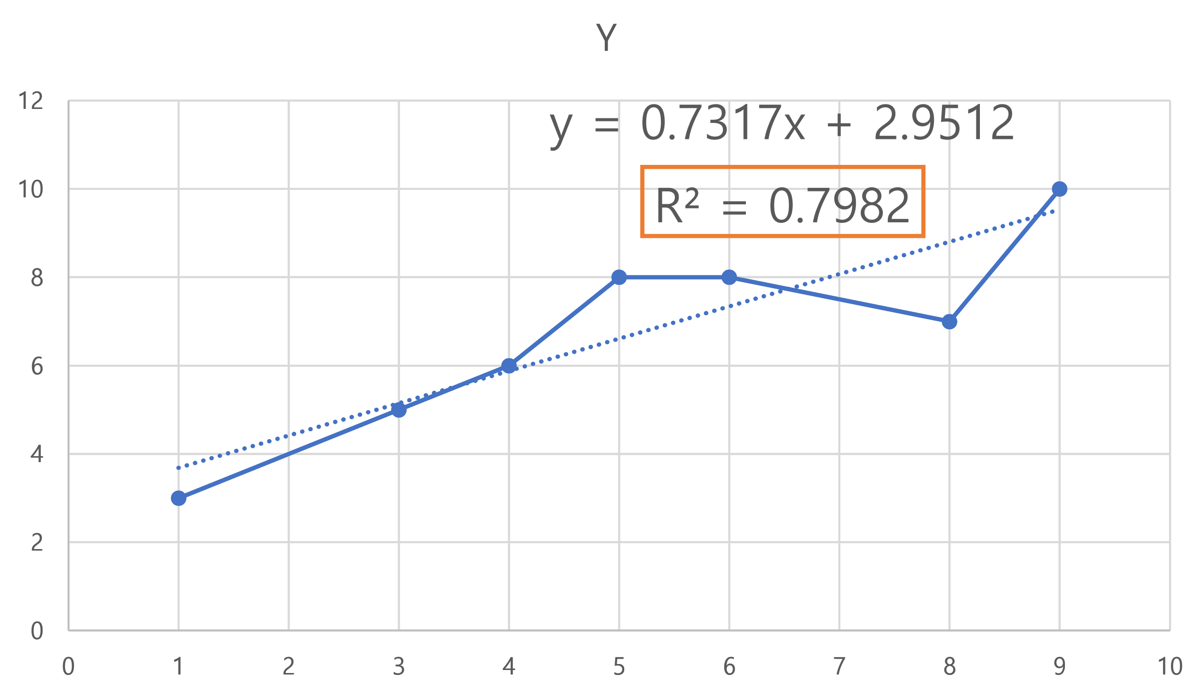
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.5791	6.7584	-2.601	0.0123
speed	3.9324	0.4155	9.464	1.49e-12



$$(dist) = -17.5791 + 3.9324 \times (speed)$$



- 지금까지 한 것: 추세선 구하기(회귀계수 추정하기)



- R-제곱 값: 직선이 데이터를 설명하는 정도, 상관계수와 관련된 값
- R-제곱 값의 정확한 의미는? 그리고 구하는 방법은?

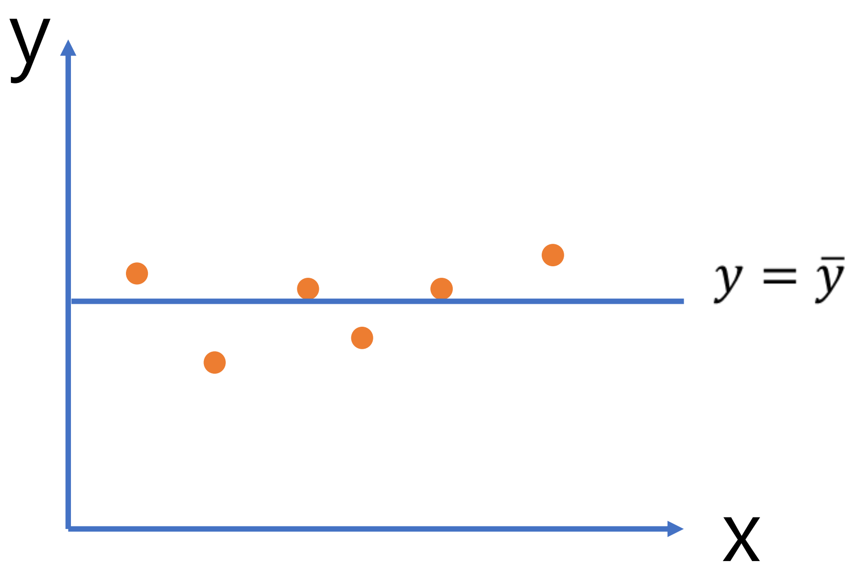
1. 회귀의 뜻과 용어
2. 회귀계수 찾는 법
3. 회귀식 해석하기
4. 모형의 적합도 평가
5. 실습
6. 잔차분석
7. 회귀분석 다시하기

2

모형의 적합도 평가

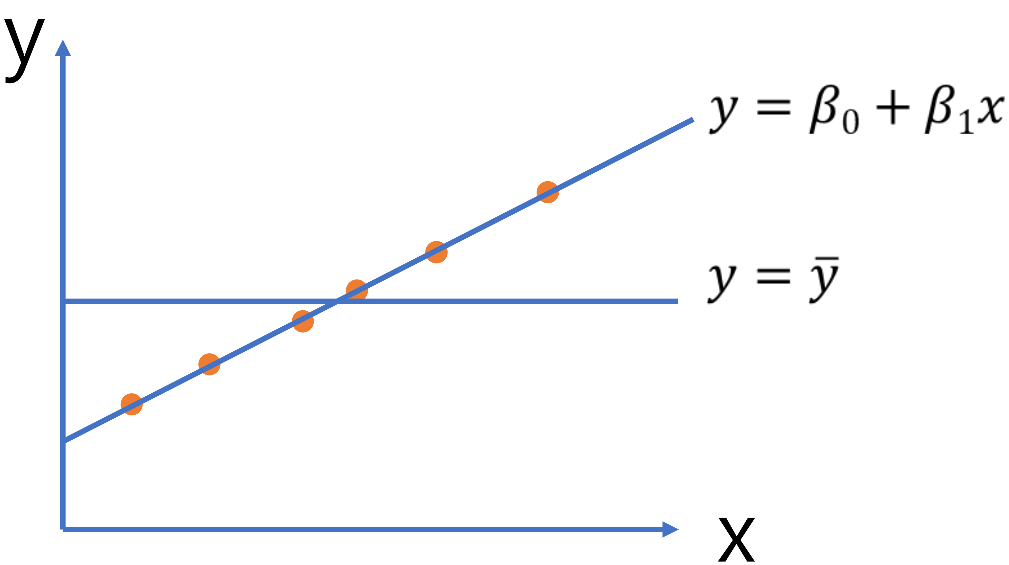
R^2

y가 x와 관련이 전혀 없는 경우, 즉 x로 y를 전혀 예측할 수 없는 경우는 오차에 의한 영향만 있기 때문에 데이터가 y의 평균 $\bar{y}$ 근처에 퍼져 있을 것이다.



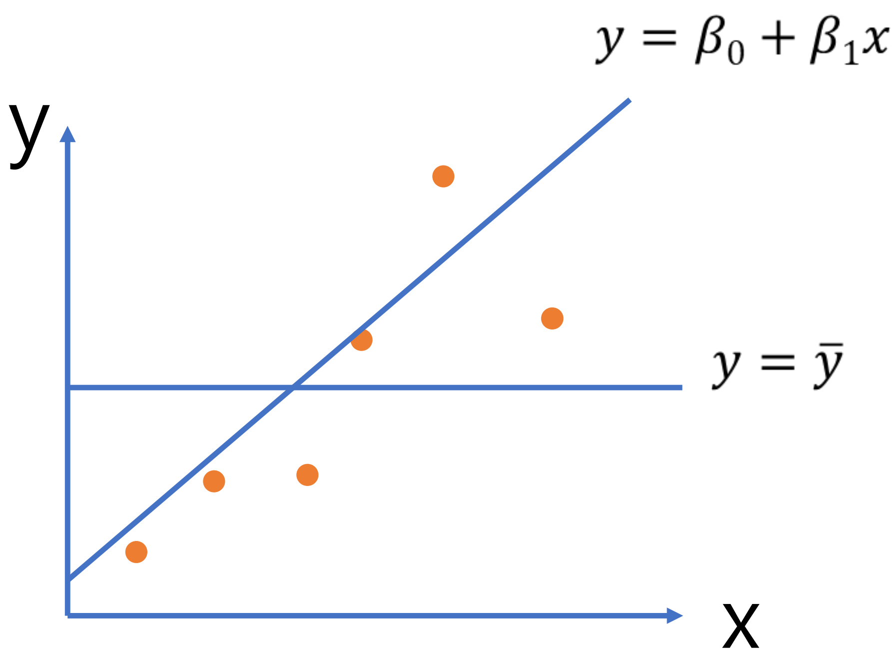
만약 x가 y와 완벽한 상관 관계가 있다면, 즉 오차가 전혀 없다면

데이터가 직선 $y = \beta_0 + \beta_1 x$ 위에 모두 존재할 것이다.



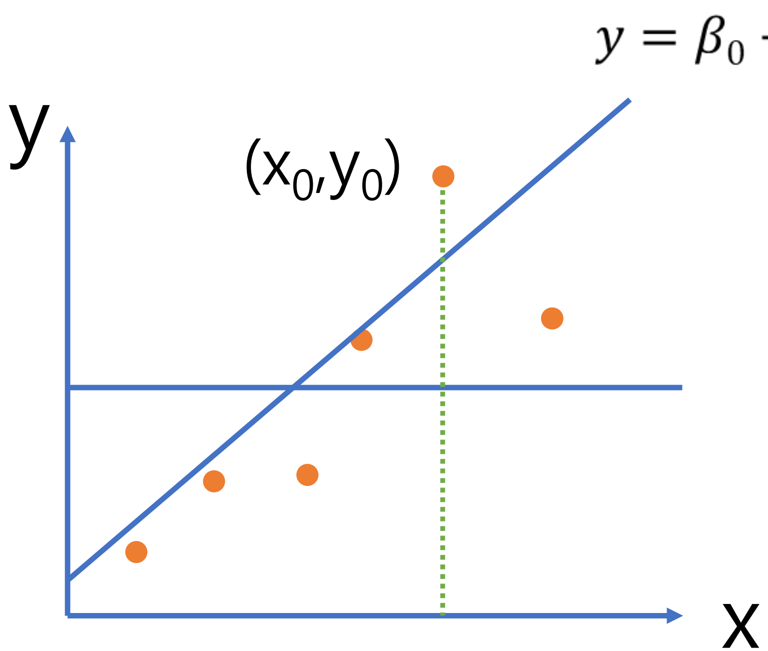
물론 실제 데이터는 두 가지 상황의 사이에 있을 것이므로,

데이터가 직선 $y = \beta_0 + \beta_1 x$ 를 중심으로 어느 정도 퍼져 있을 것이다.



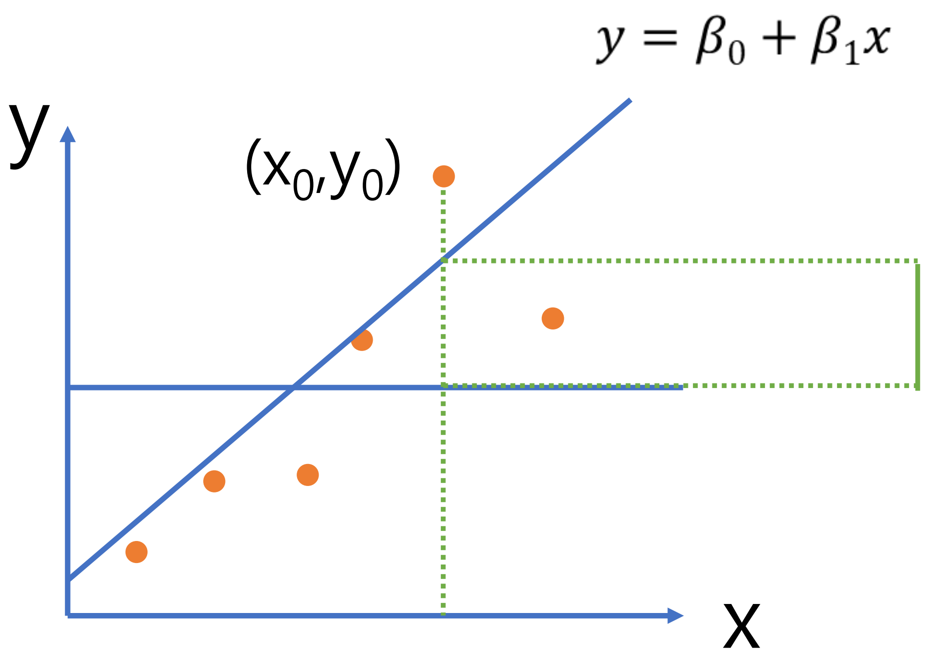
따라서 어떤 점 (x_0,y_0)에 대해 $\bar{y}$와 $\beta_0 + \beta_1 x_0$의 차이를 고려한다면,

직선 $y = \beta_0 + \beta_1 x$가 주어진 점 (x_0,y_0)를 얼마나 잘 설명하고 있는지 알 수 있다.

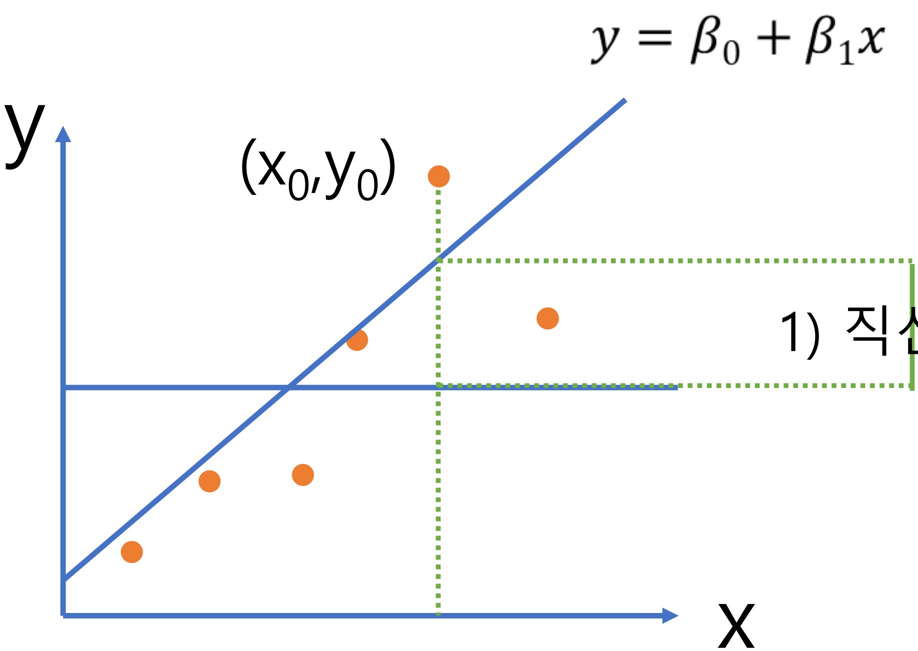


따라서 어떤 점 (x_0, y_0)에 대해 $\bar{y}$와 $\beta_0 + \beta_1 x_0$의 차이를 고려한다면

직선 $y = \beta_0 + \beta_1 x$가 주어진 점 (x_0, y_0)를 얼마나 잘 설명하고 있는지 알 수 있다.



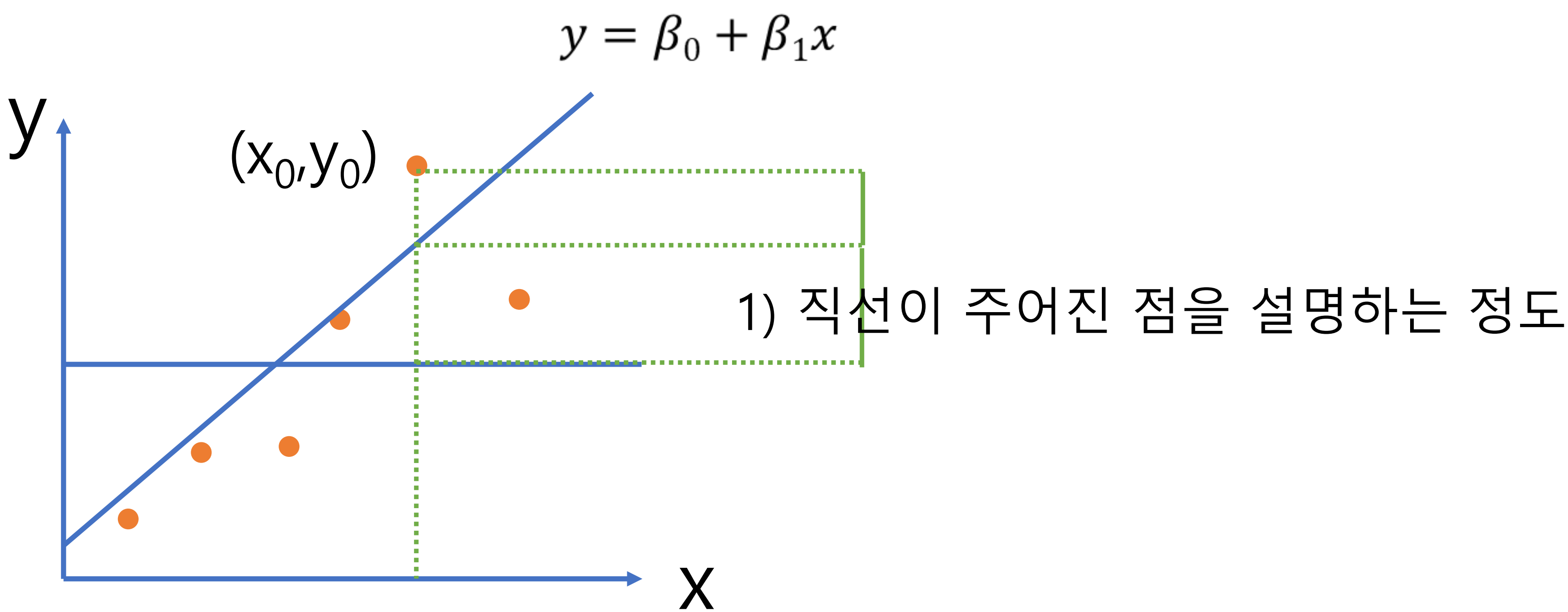
따라서 어떤 점 (x_0,y_0)에 대해 $\bar{y}$와 $\beta_0+\beta_1 x_0$의 차이를 고려한다면(1) 표시),
직선 $y=\beta_0+\beta_1 x$가 주어진 점 (x_0,y_0)를 얼마나 잘 설명하고 있는지 알 수 있다.



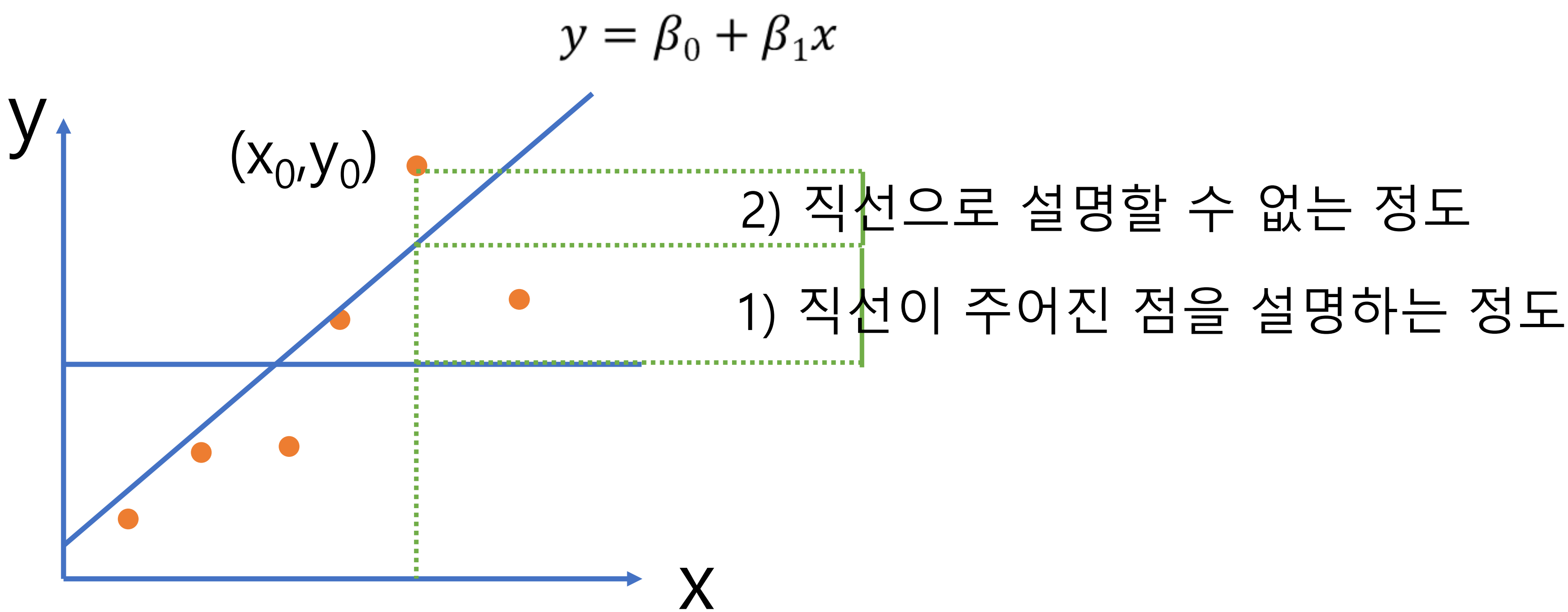
1) 직선이 주어진 점을 설명하는 정도

여기서 실제 값 y_0와 $\beta_0 + \beta_1 x_0$는 값이 다를 수 있다.

이러한 차이는 직선으로 설명할 수 없는, 오차 ε에 해당하는 부분이다.

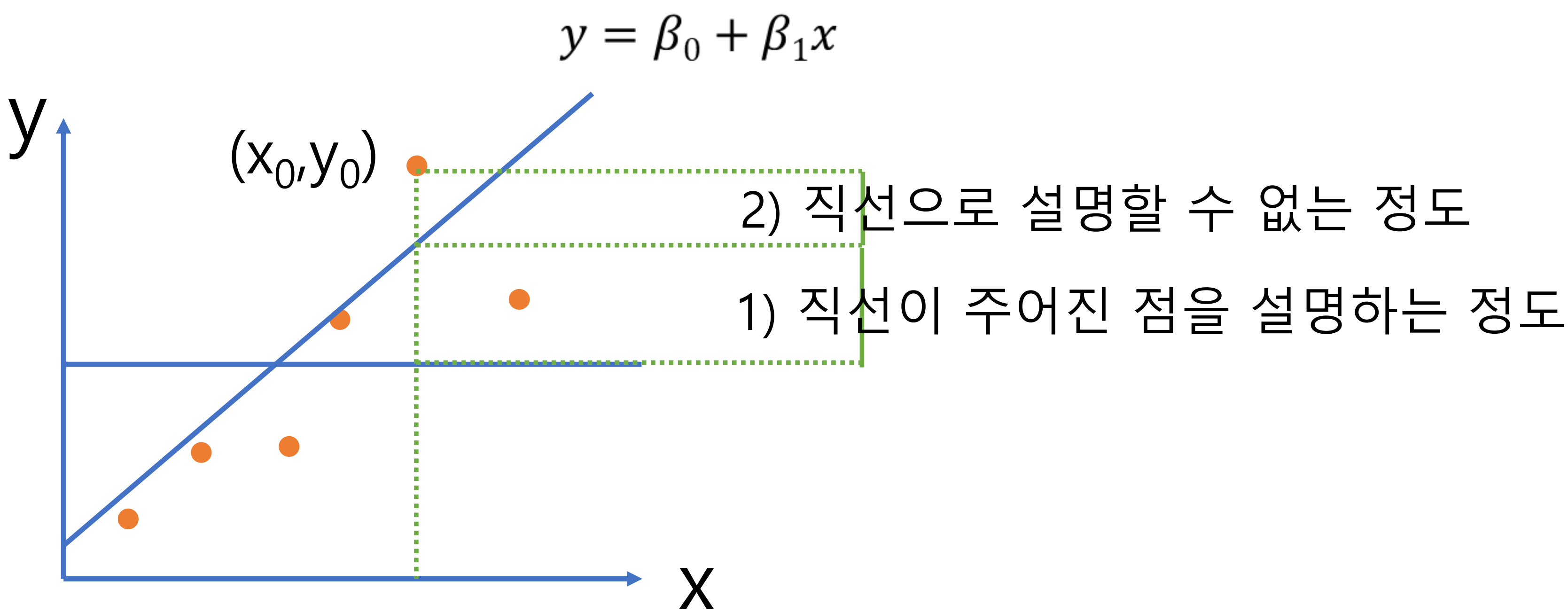


여기서 실제 값 y_0와 $\beta_0 + \beta_1 x_0$는 값이 다를 수 있다(2) 표시).
이러한 차이는 직선으로 설명할 수 없는, 오차 ε에 해당하는 부분이다.



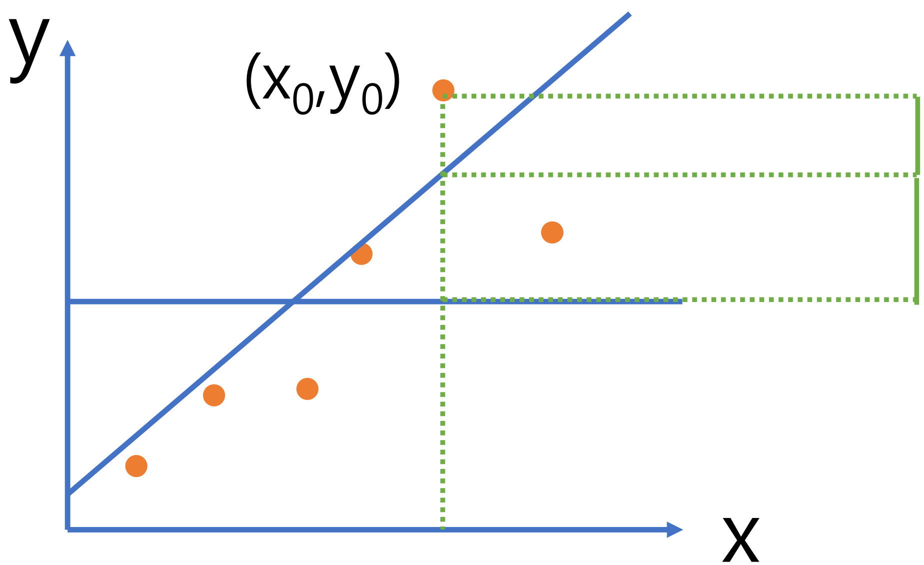
여기서 $1) + 2) = (y_0 - \beta_0 + \beta_1 x_0) + (\beta_0 + \beta_1 x_0 - \bar{y}) = y_0 - \bar{y}$

= 점 (x_0, y_0)의 변동성



한 가지 문제점은 β_0 , β_1과 오차 ε은 추상적인 개념으로 관측할 수 없다.

$$y = \beta_0 + \beta_1 x$$

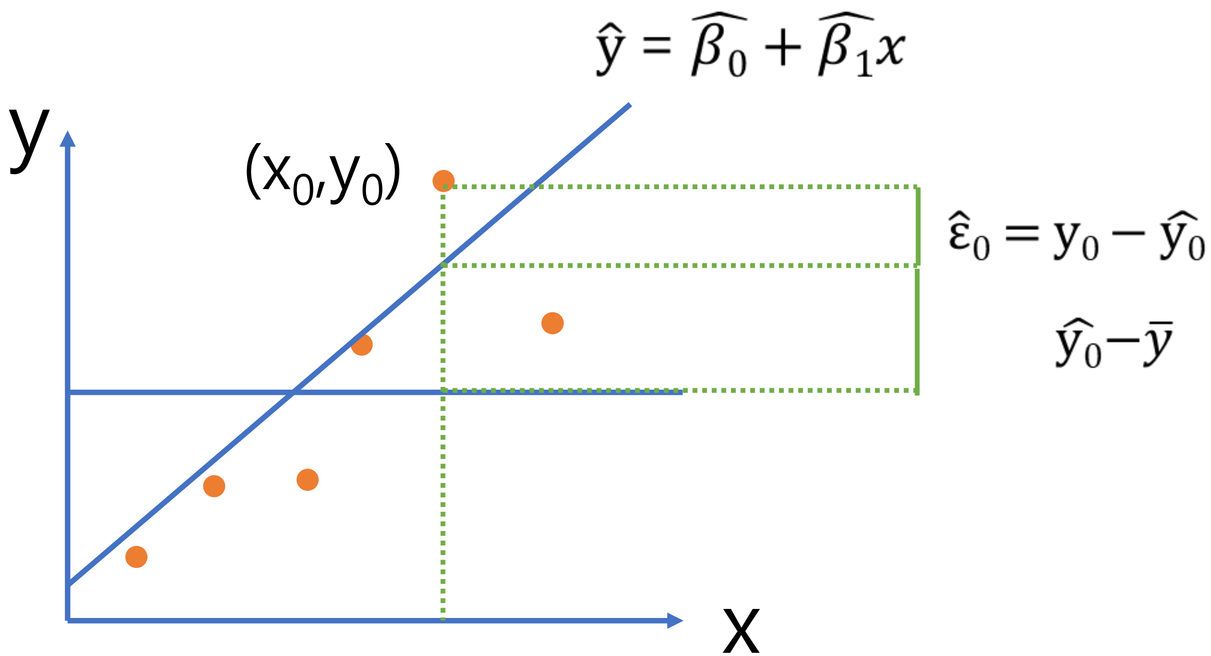


$$\varepsilon_0 = y - (\beta_0 + \beta_1 x_0)$$

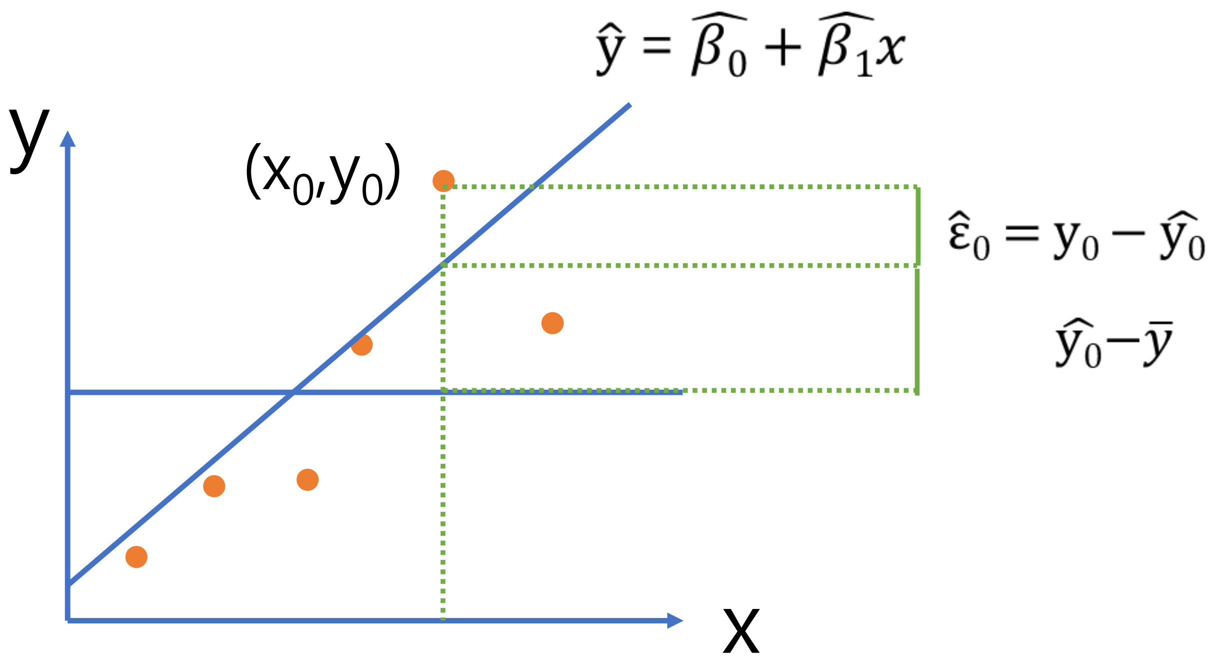
$$\beta_0 + \beta_1 x_0 - \bar{y}$$

한 가지 문제점은 β_0, β_1과 오차 ε은 추상적인 개념으로 관측할 수 없다.

-> 추정값 $\widehat{\beta_0}$, $\widehat{\beta_1}$을 사용하고, 오차 ε 대신 잔차 $\hat{\varepsilon} = \hat{y} - y$를 정의하여 사용한다.



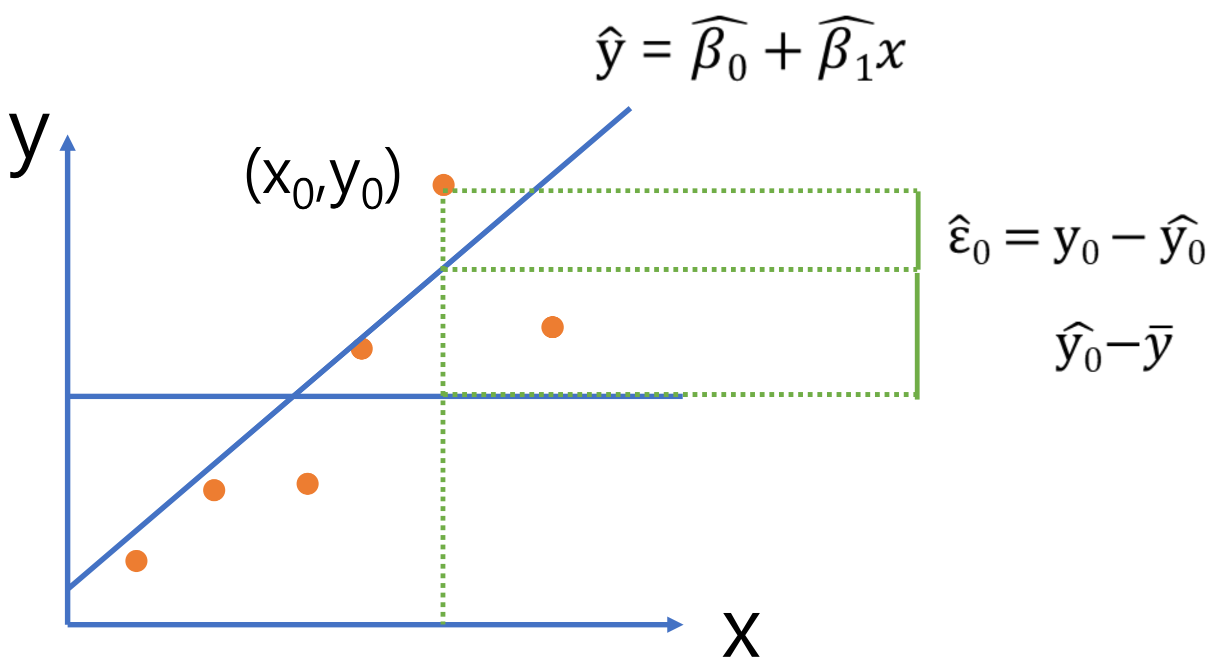
$$y_0 - \bar{y} \quad = \quad \widehat{y_0} - \bar{y} \quad + \quad \hat{\varepsilon}_0$$



Goal 3. 단순 선형회귀분석 : 모형의 적합도 평가

$$y_0 - \bar{y} \quad = \quad \widehat{y_0} - \bar{y} \quad + \quad \hat{\varepsilon}_0$$

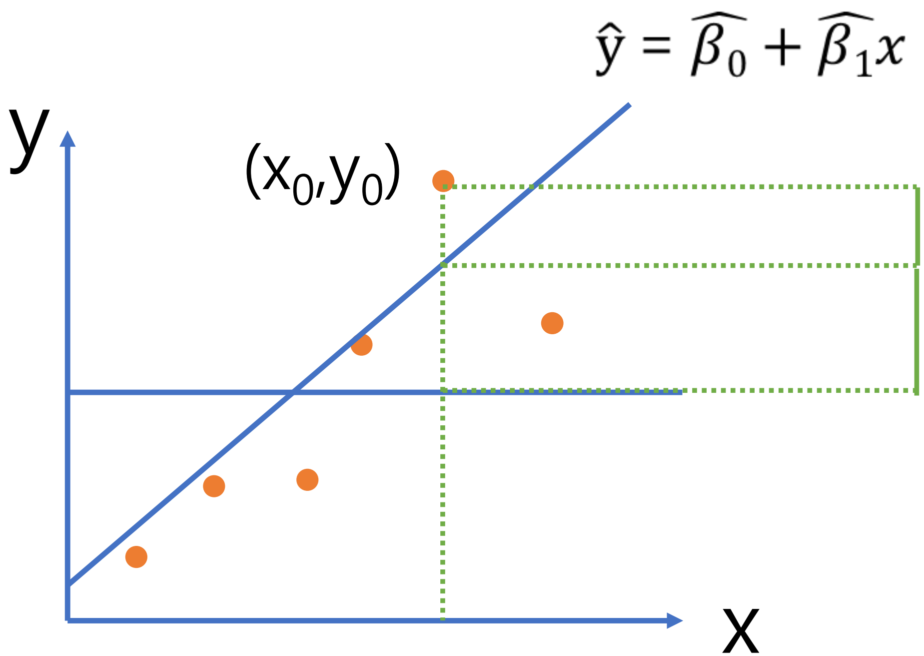
(전체 변동성) = (직선으로 설명되는 변동성) + (직선으로 설명되지 않는 변동성)



단순 선형회귀분석: 모형의 적합도 평가

$$y_0 - \bar{y} \quad = \quad \widehat{y_0} - \bar{y} \quad + \quad \hat{\varepsilon}_0$$

(전체 변동성) = (직선으로 설명되는 변동성) + (직선으로 설명되지 않는 변동성)



$\hat{\varepsilon}_0 = y_0 - \widehat{y_0}$ <- 잔차 제곱합 (E)

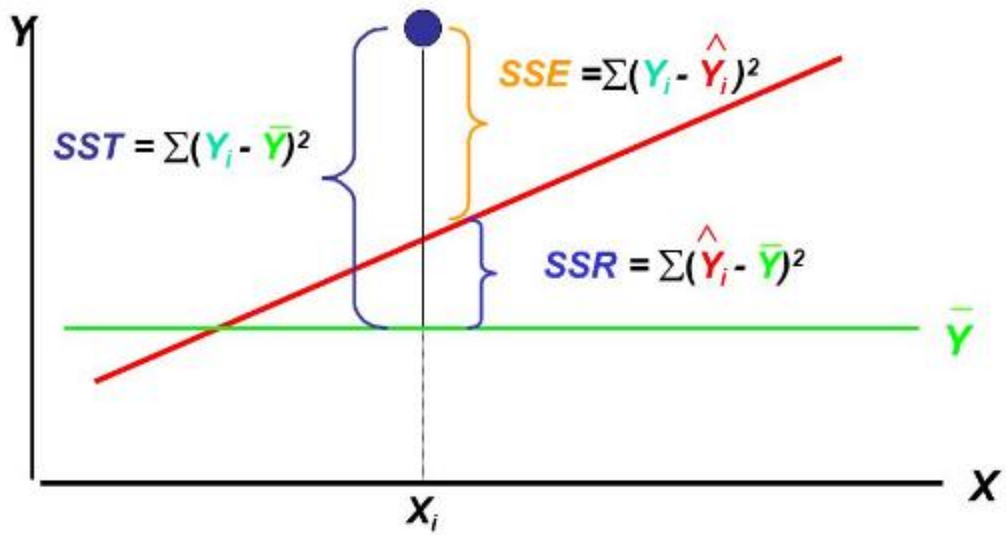
$\widehat{y_0} - \bar{y}$ <- 회귀 제곱합 (R)

Goal 3. 단순 선형회귀분석 : 모형의 적합도 평가

Sum of Squares Decomposition

SST: 총 제곱합, SSR: 회귀제곱합, SSE: 잔차제곱합

실제로 제곱합을 이용해 여러 추정 및 검정이 이루어진다.



$$\sum_{j=1}^{n}(y_j-\bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j-\bar{y})^2 + \sum_{j=1}^{n}\hat{\varepsilon}_j^{\,2}$$

SST SSR SSE

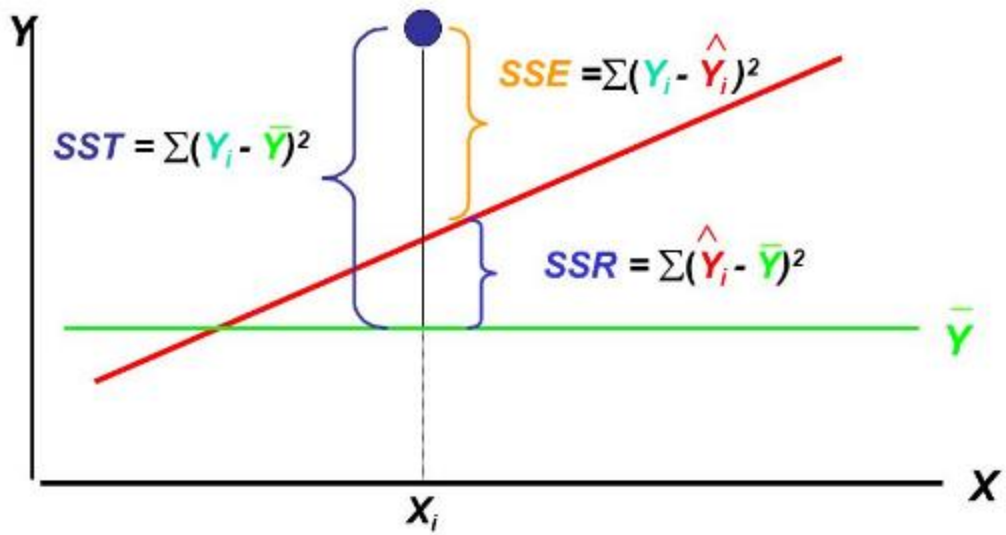
SST=SSR+SSE의 관계가 항상 성립한다.

Goal 3. 단순 선형회귀분석 : 모형의 적합도 평가

SST: 전체 데이터의 변동성(y값이 평균과 떨어져있는 정도)

SSR: 회귀직선으로 설명되는 변동성

SSE: 회귀직선으로 설명되지 않는 변동성($\hat{\varepsilon}_j = \hat{y}_j - y_j$; 잔차)



$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}\hat{\varepsilon}_j^{\,2}$$

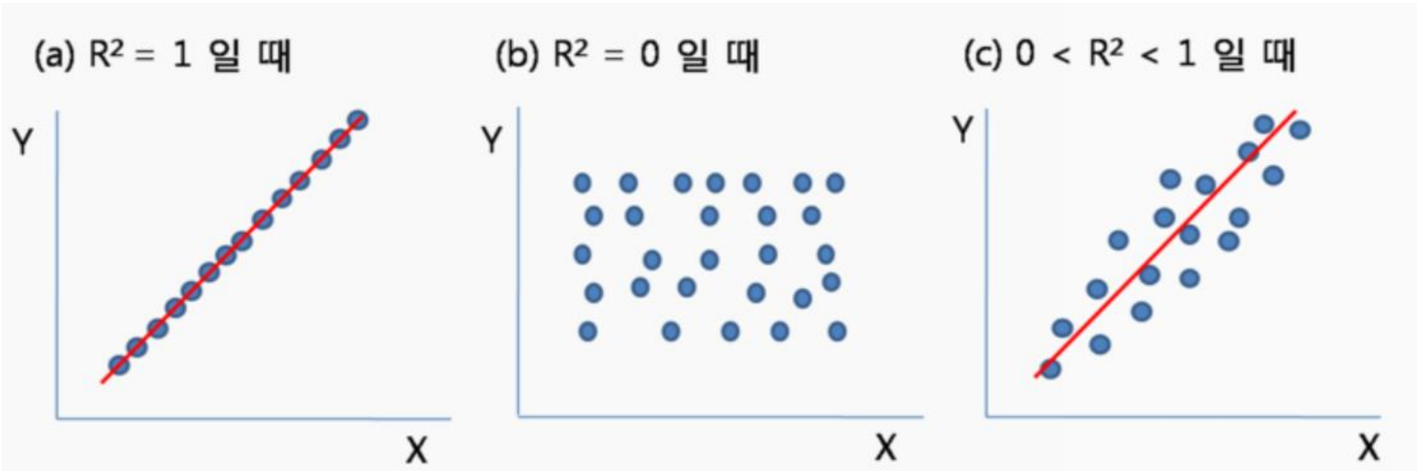
SST
(전체 데이터의 변동성)

SSR
(회귀식에 의해 설명되는 변동성)

SSE
(회귀식으로 설명할 수 없는 변동성)

R^2(결정계수): 반응변수(y)의 전체 변동 중 설명변수(x)가 차지하는 변동의 비율

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \leq R^2 \leq 1 (\because SST=SSR+SSE)$$



설명변수 없이 항상 $\bar{y}$를 예측값으로 사용하는 모형과 회귀모형의 성능을 비교

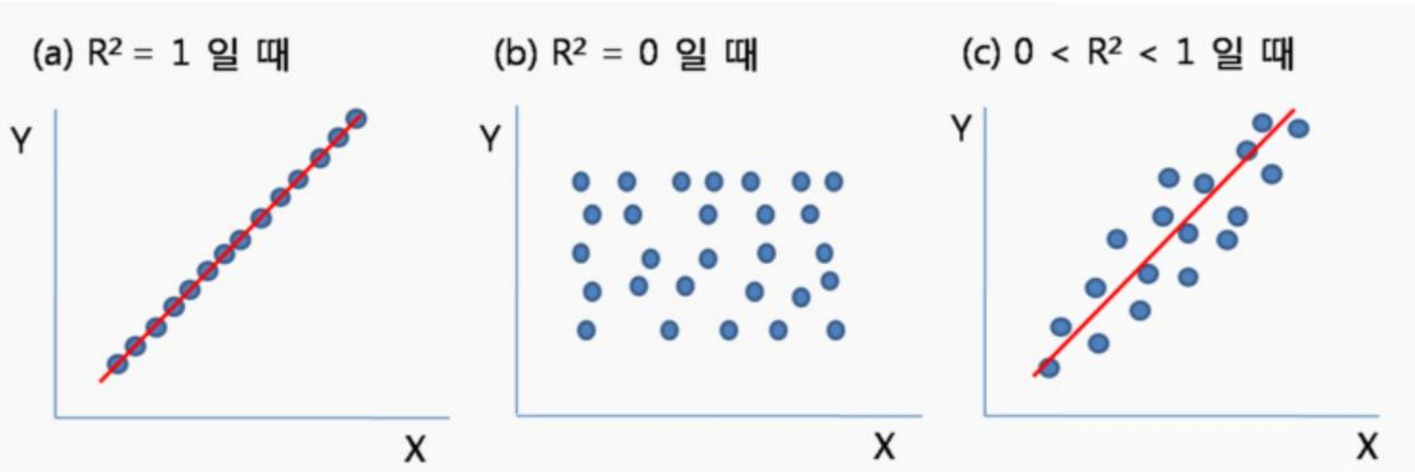
R^2 가 클수록 모형이 데이터를 잘 설명한다.

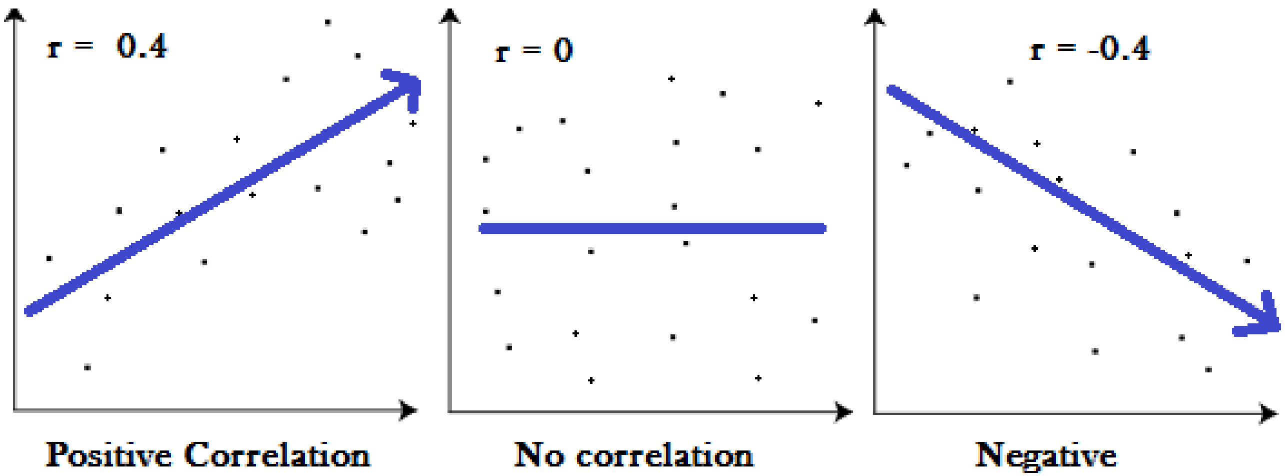
R^2 =1 : 모든 측정값들이 회귀직선 위에 있는경우

R^2 =0 : 추정된 회귀직선은X와 Y의 관계를 전혀 설명하지 못함

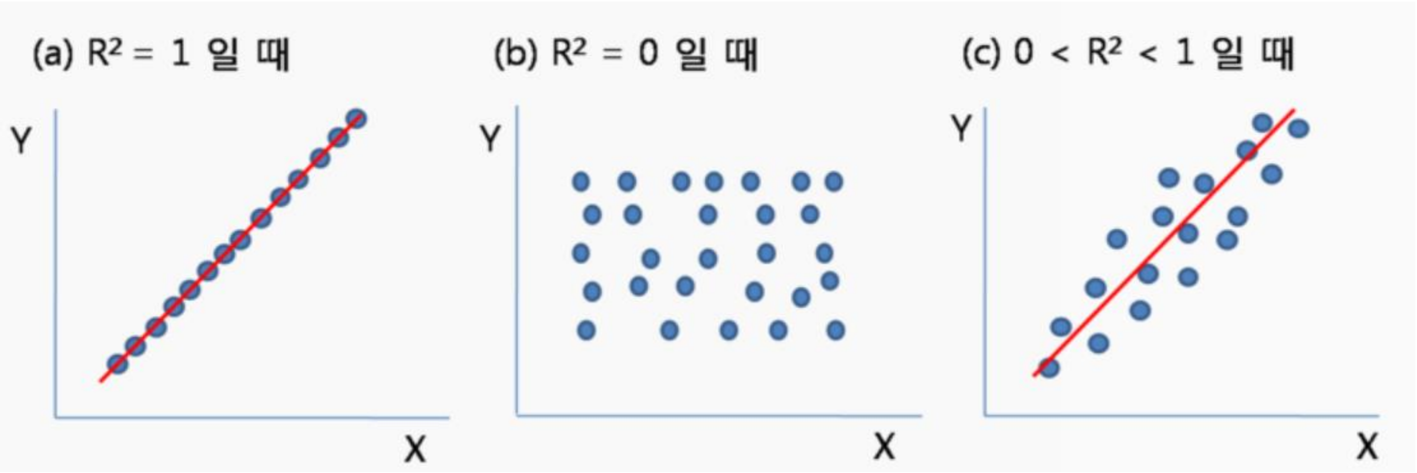
Goal 3. 단순 선형회귀분석 : 모형의 적합도 평가

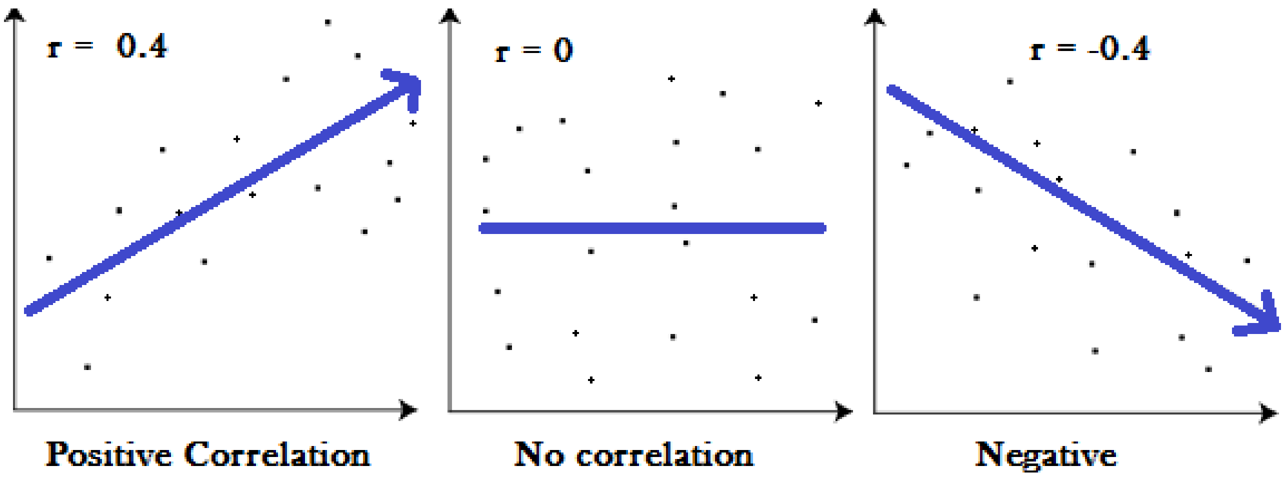
(참고) 설명변수가 하나인 경우 결정계수와 상관계수의 관계?





(참고) 표본상관계수를 r이라 할 때, $R^2 = r^2$





1. 회귀의 뜻과 용어
2. 회귀계수 찾는 법
3. 회귀식 해석하기
4. 모형의 적합도 평가
5. 실습
6. 잔차분석
7. 회귀분석 다시하기

3

실 습

lm(y ~ x, data=   )

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

데이터 살펴보기

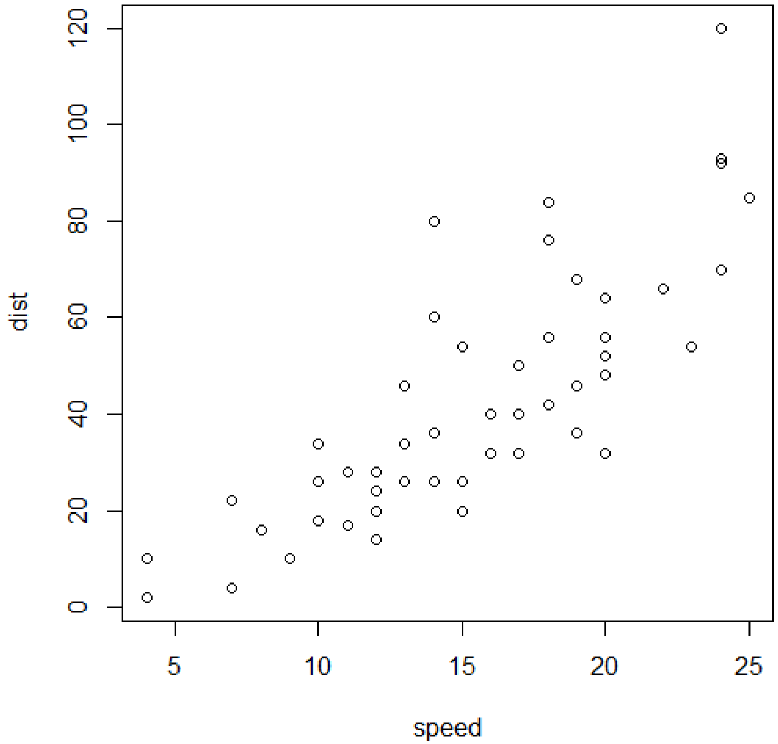
```
> data1 <- cars
> head(data1, 10)
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
```

“cars” 데이터(내장 데이터) : 브레이크가 작동되는 순간의 자동차의

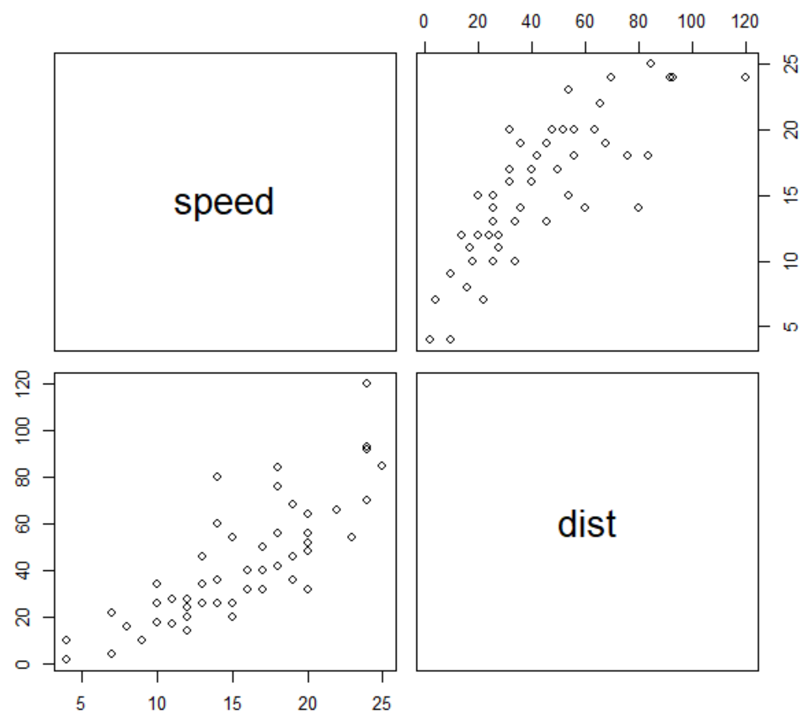
주행 속도(Speed)에 따른 자동차 제동 거리(Dist)를 조사한 자료

데이터 살펴보기

```
> plot(data1)
```

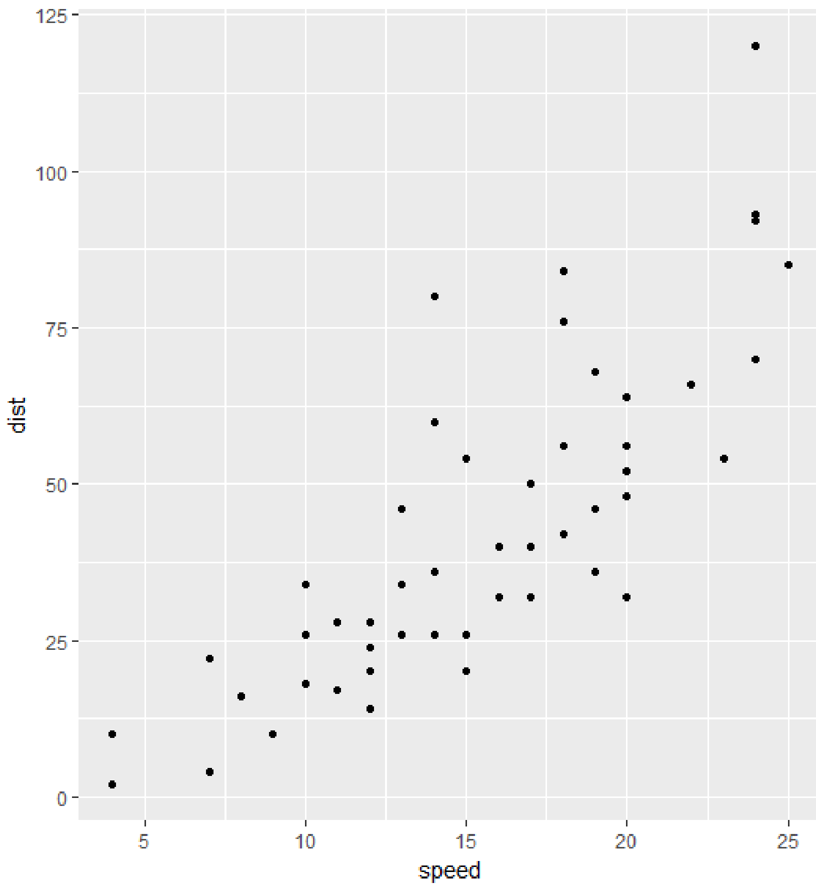



```
> pairs(data1)
```



데이터 살펴보기

```r
library(ggplot2)
ggplot(data=data1, aes(x=speed, y=dist)) + geom_point()
```



(참고)
패키지 "ggplot2"
데이터 시각화에 쓰이는 대표적인 패키지

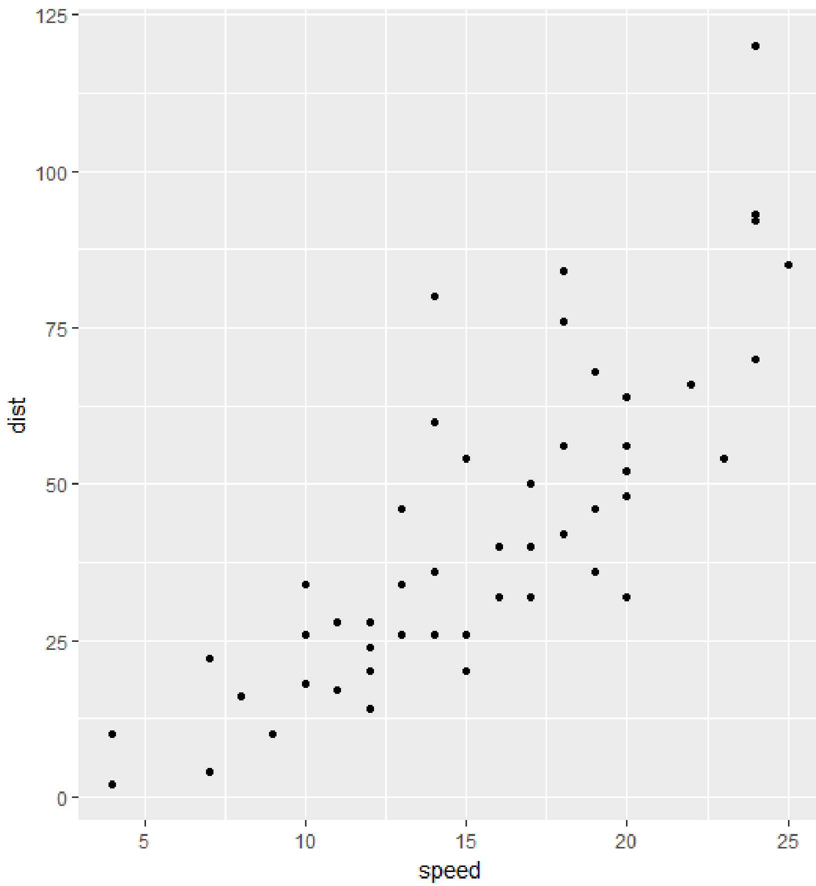
CRAN에 등록된 패키지는 먼저 명령어 install.packages("@@")를 이용해 설치한 뒤 이용 가능

```r
> install.packages("ggplot2")
```

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

데이터 살펴보기

```
library(ggplot2)
ggplot(data=data1, aes(x=speed, y=dist)) + geom_point()
```



산점도를 확인한 결과 둘 사이에는 양의 상관관계가(선형적 연관성이) 존재하는 것으로 짐작할 수 있다.

-> 상관분석을 통해 상관관계의 유무를 확인해보자.

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

상관분석

```
> cor(data1)
          speed      dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
```

cor( ): 상관행렬(correlation matrix)

```
> cor.test(data1$speed, data1$dist)

        Pearson's product-moment correlation

data:  data1$speed and data1$dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```

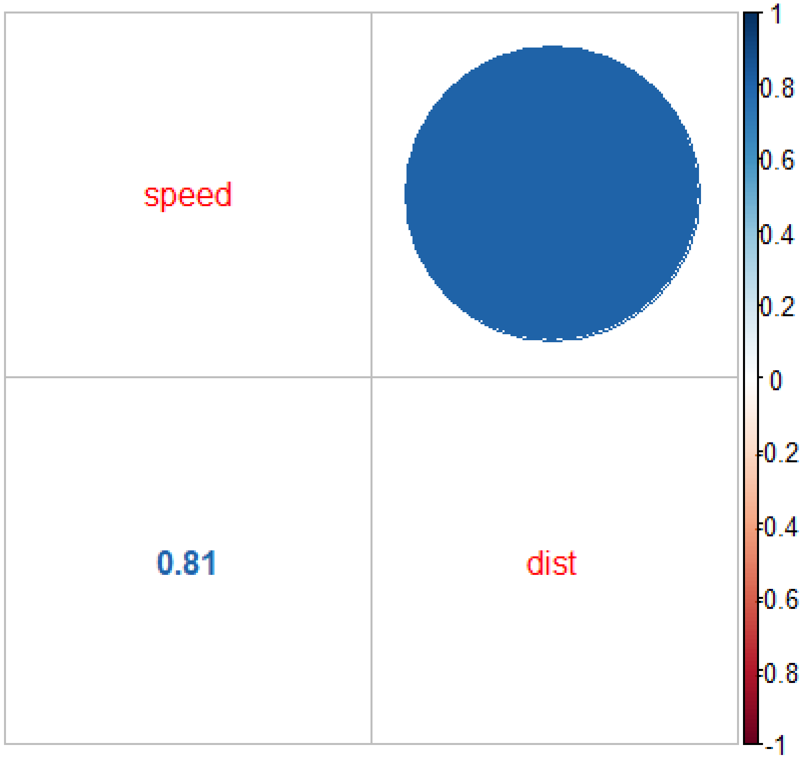
cor.test(x, y): x, y의 상관관계 검정

p-value가 값이 매우 작으므로 유의,
추정된 상관계수 값은 약 0.8069

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

상관분석

```
> #install.packages("corrplot")
> corrplot::corrplot.mixed(cor(data1))
```



: 주석 표시
뒤에 오는 말은 R에서 무시한다
(실행되지 않는다).

패키지 "corrplot"

상관행렬을 시각화하는데 사용

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

모형 적합

```
> fit.cars <- lm(dist ~ speed, data=cars)
> fit.cars

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932

> summary(fit.cars )
```

회귀모형은 lm(y~x) 함수를 이용한다. 여기서 lm은 선형 모형(Linear Model)의 약자로, 회귀분석, 분산분석 등이 대표적인 선형모형이다.

정답은 $y = -17.579 + 3.932*x$ 였던 것이었다!

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

Summary( )는 3가지 정보를 보여준다

```
> fit.cars<-lm(dist~speed, data=cars)
> summary(fit.cars)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

믿을만 한가?

```
> fit.cars<-lm(dist~speed, data=cars)
> summary(fit.cars)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

- 귀무가설은 'H_0: 계수(또는 절편)이 0'이다.
- 대립가설은 'H_1: 계수(또는 절편)이 0이 아님'이다.

(참고) 단순회귀분석에서는

- 변수 speed의 유의확률과 회귀직선 전체의 유의확률이 같음
- $R^2 = 0.6511 = (0.8069)^2 = r^2$가 성립

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

사실은 12가지 정보

```
> str(fit.cars)
List of 12
```

선형회귀한 object는 list 형이며 12개의 열로 이루어져 있다.

각각의 열에 들어있는 정보를 살펴보자

```
> names(fit.cars)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

> attach(fit.cars)
> coefficients
(Intercept)       speed
 -17.579095    3.932409

> residuals
        1         2         3         4         5
 3.849460 11.849460 -5.947766 12.052234  2.119825
```

```
resid(fit.cars)
residuals(fit.cars)
residuals.lm(fit.cars)
```

잔차들

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

```
> effects
 (Intercept)        speed
-303.9144946  145.5522550   -8.1154395    9.8845605

   0.1941147   -9.4963311   -5.1867770    2.8132230
```

```
> rank
[1] 2
```

```
> fitted.values
        1          2          3          4         5
-1.849460 -1.849460  9.947766  9.947766 13.880175
```

회귀식에 대입한 적합값들

```
> df.residual
[1] 48
```

잔차의 자유도 : 50 obs에서 온 것

```
> call
lm(formula = dist ~ speed, data = cars)
```

선형회귀함수

```
> model
   dist speed
1     2     4
2    10     4
```

초기 입력값들

fitted.values 값은 이렇게 나온 것

```
> fitted.values
        1         2         3         4         5 
-1.849460 -1.849460  9.947766  9.947766 13.880175 
```

회귀식에 대입한 적합값들

```
> fit.cars$coefficients
(Intercept)       speed 
 -17.579095    3.932409 
```

```
f1 <- function(x){
  y = -17.579095 + 3.932409*x
  return(y)
}

x <- c(cars$speed)
f1(x)
```

```
> f1(x)
 [1] -1.849459 -1.849459  9.947768  9.947768 13.880177
 [6] 17.812586 21.744995 21.744995 21.744995 25.677404
```

Goal 4. 예제 R 코드를 통해 단순 선형회귀분석을 실습해보자

residuals 값은 이렇게 나온 것

```
> residuals
        1         2         3         4         5 
 3.849460 11.849460 -5.947766 12.052234  2.119825 
```

잔차들

```
f2 <- function(x, yi){
  y_hat = -17.579095 + 3.932409*x
  residuals = yi - y_hat
  return(residuals)
}

x <- c(cars$speed)
yi <- c(cars$dist)
f2(x, yi)
```

```
> f2(x, yi)
 [1]   3.849459  11.849459  -5.947768  12.052232   2.119823
 [6]  -7.812586  -3.744995   4.255005  12.255005  -8.677404
```

summary 와 비교해 보자

```
> summary(fit.cars$residuals)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-29.069   -9.525   -2.272    0.000    9.215   43.201 
```

예측하기

```
predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
        interval = c("none", "confidence", "prediction"),
        level = 0.95, type = c("response", "terms"),
        terms = NULL, na.action = na.pass,
        pred.var = res.var/weights, weights = 1, ...)
```

Arguments

`object`	Object of class inheriting from `"lm"`
`newdata`	An optional data frame in which to look for variables with which to predict.

```
> new <- data.frame(speed =c(122))
> predict(fit.cars, newdata = new)
       1
462.1748
```

예측하기

```
> new <- data.frame(speed =c(122, 125, 130, 133))
> predict(fit.cars, newdata = new)
       1        2        3        4 
462.1748 473.9720 493.6340 505.4313 
```

```
> new <- data.frame(speed =c(122))
> predict(fit.cars, newdata = new, interval = "confidence")

       fit      lwr      upr
1 462.1748 373.0091 551.3405
```

```
> new <- data.frame(speed =c(122, 125, 130, 133))
> predict(fit.cars, newdata = new, interval = "confidence")
       fit      lwr      upr
1 462.1748 373.0091 551.3405
2 473.9720 382.3029 565.6411
3 493.6340 397.7923 589.4758
4 505.4313 407.0857 603.7768
```

예측하기

```
> new <- data.frame(speed =c(122))
> predict(fit.cars, newdata = new, interval = "confidence")
       fit      lwr      upr
1 462.1748 373.0091 551.3405
> predict(fit.cars, newdata = new, interval = "confidence", level = 0.9)
       fit      lwr      upr
1 462.1748 387.7949 536.5547
```

```
> predict(fit.cars, newdata = new, interval = "prediction")
       fit      lwr      upr
1 462.1748 367.7993 556.5503
```