강원대학교
AI 소프트웨어학과

머신러닝2

- 단순 회귀분석 -

KNU 강원대학교

기계학습

추론통계

- 목적: 추론 통계는 주로 샘플을 기반으로 모집단에 대한 결론을 도출하는 것을 목표(예를 들어 평균 또는 표준 편차와 같은 모집단 매개변수를 추정하거나 해당 매개변수에 대한 가설을 테스트하는데 사용 가능)
- 방법: 이는 종종 신뢰 구간을 구성하거나 가설 테스트를 수행하여 수행됨
- 결과: 추론 통계는 종종 불확실성을 정량화 하여 처리함(예: "우리는 모집단 평균이 이 구간에 있다고 95% 확신함)

기계 학습

- 목적: 기계 학습의 주요 목표는 예측을 하거나 새로운 보이지 않는 데이터를 분류할 수 있는 모델을 개발하는 것
- 표본 집단을 사용하지만 모집단의 속성을 추론하기보다는 새로운 데이터의 일반화에 초점
- 방법: 표본 집단에서 모델을 훈련시킨 다음 별도의 세트(테스트 세트)에서 성능을 평가함
- 제한 사항: 과대적합은 기계 학습의 주요 고려 대상이고, 모델이 노이즈 및 이상치(값)을 포함하여 훈련 데이터를 너무 잘 학습하여 새 데이터에 대한 성능을 손상시키는 경우에 발생함

기계학습

추론통계와 기계학습 모두 샘플 데이터를 사용하지만 목표는 다음과 같이 다름

- 추론통계 : 샘플이 있고 더 넓은 모집단을 이해하려고 함
- 기계학습 : 샘플이 있고 새로운 데이터에 대해 정확한 예측 함
- 데이터 과학에서는 전통적인 통계와 기계 학습 사이의 경계가 흐려질 수 있음

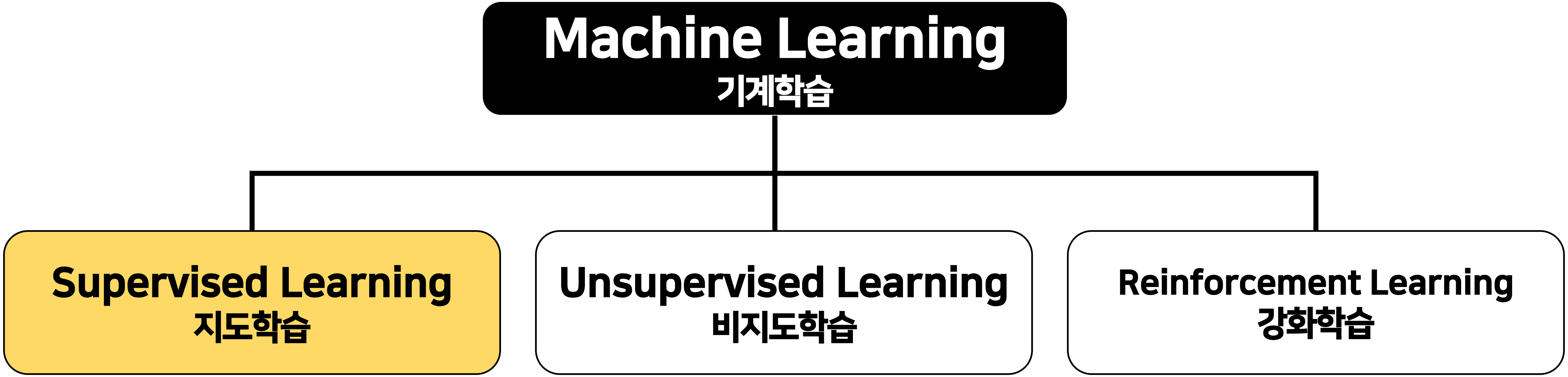
추론통계

- 분산분석 → 개 이상의 그룹의 평균을 비교하여 적어도 하나의 그룹 평균이 다른 그룹과 유의미하게 다른지 여부를 결정함
- 주로 범주형 독립 변수에 초점을 맞춤

기계학습

- 종속 변수와 하나 이상의 독립 변수 간의 관계를 모델링함
- 이러한 관계의 강도와 방향을 정량화하고 예측에 사용될 수 있음

기계학습의 종류

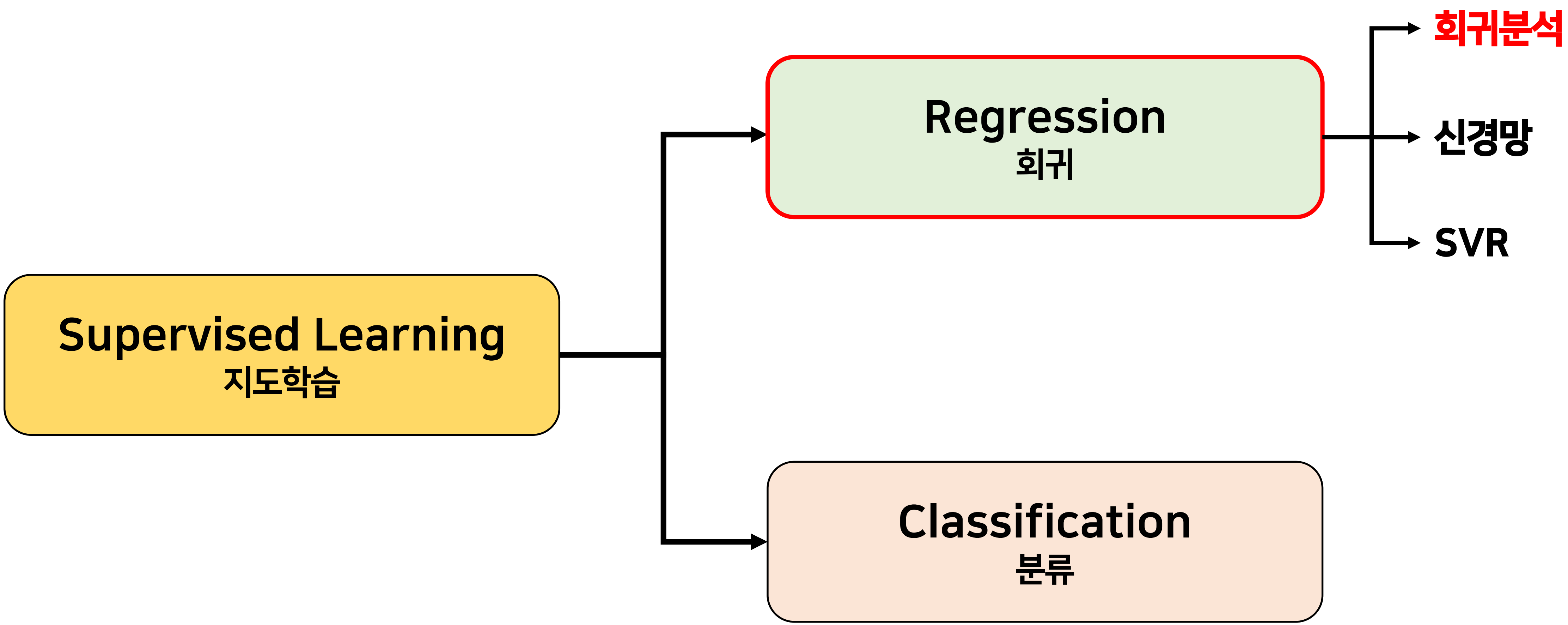


정답이 있는 데이터 → Label 존재
데이터 분류 / 정확한 결과 예측

정답이 없는 데이터 → Label 없음
데이터의 패턴 / 구조를 통해 분류

행동에 대한 보상을 수여함
누적 보상을 최적화 하는 의사결정

지도학습


Supervised Learning
지도학습
Regression
회귀
Classification
분류
회귀분석
신경망
SVR

회귀분석(Regression Analysis)

변수(variable)는 개체의 어떤 특징을 나타내는 것

사람이 키가 클수록 몸무게가 커지는 현상에 대해 다음과 같은 질문을 할 수 있음

(1) 키와 몸무게는 서로 관련이 있는가?

(2) 관련이 있다면 키와 몸무게의 관계를 수학적 함수로 나타낼 수 있는가?

(3) 수학적 함수를 이용하여 키로부터 몸무게를 예측할 수 있는가?

회귀분석(Regression Analysis)

회귀분석의 구조(독립변수와 종속변수)

변수란? 변수는 주어진 상황에서 다양하거나 다른 값을 가질 수 있는 특성, 속성 또는 수량을 나타냄

종속변수 : 하나 이상의 독립변수의 변화에 의해 변화가 설명되거나 예측되는 연속형 변수

독립변수 : 예측 변수 또는 설명 변수라고도 하는 독립 변수는 종속 변수에 영향을 미칠 것으로 가정되는 연속형 변수

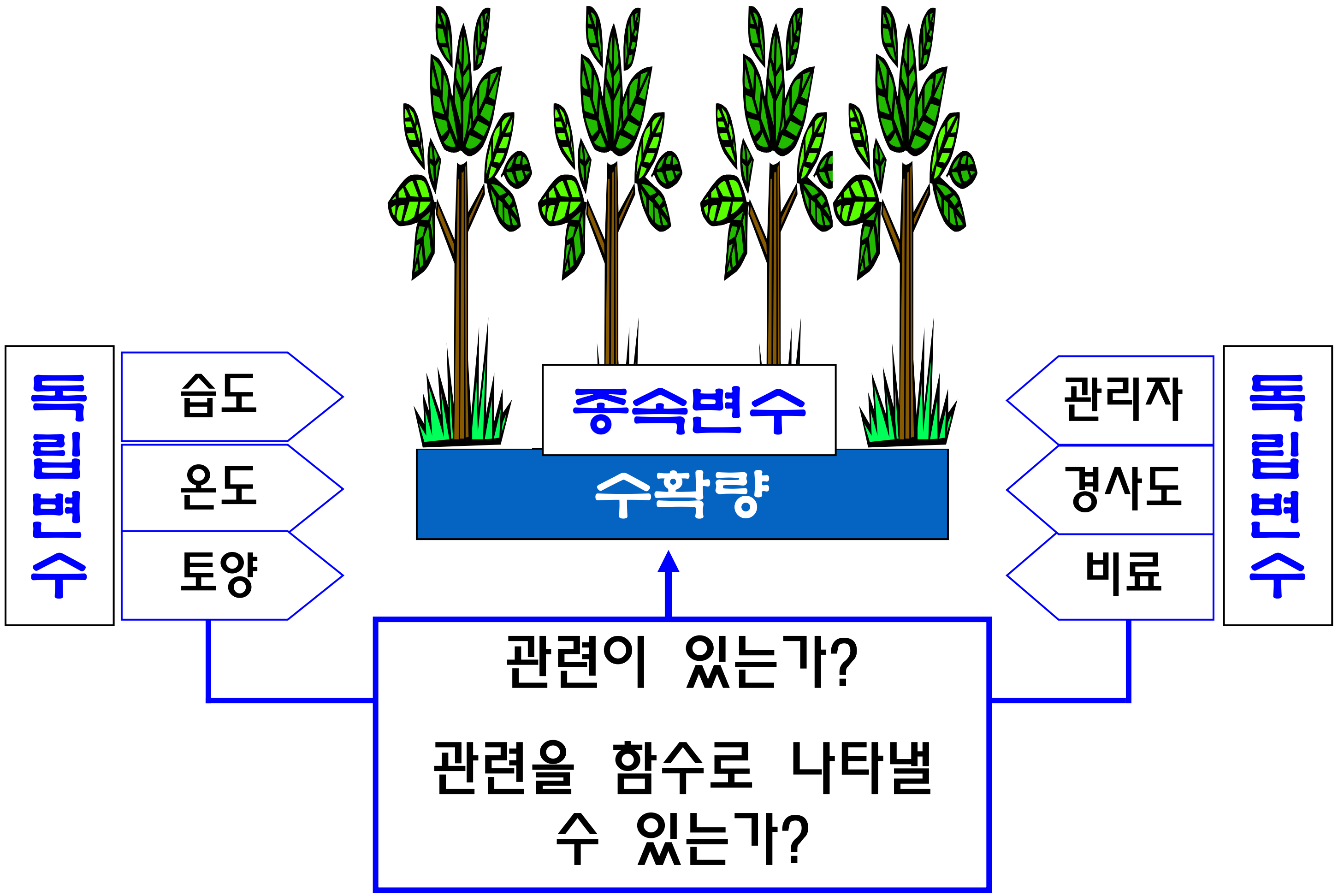
02 회귀(Regression)

회귀분석(Regression Analysis)

변수(variable)는 개체의 어떤 특징을 나타내는 것

- 변수 (variable)
  - 양적 변수 (quantitative variable)
    - 연속형 변수(continuous variable): 키, 몸무게...
    - 이산형 변수(discrete variable): 성공횟수, 교통사고건수...
  - 질적 변수 (qualitative variable)
    - 순위형 변수(ordinal variable): 학년, 선호도...
    - 명목형 변수(nominal variable): 성별, 혈액형...

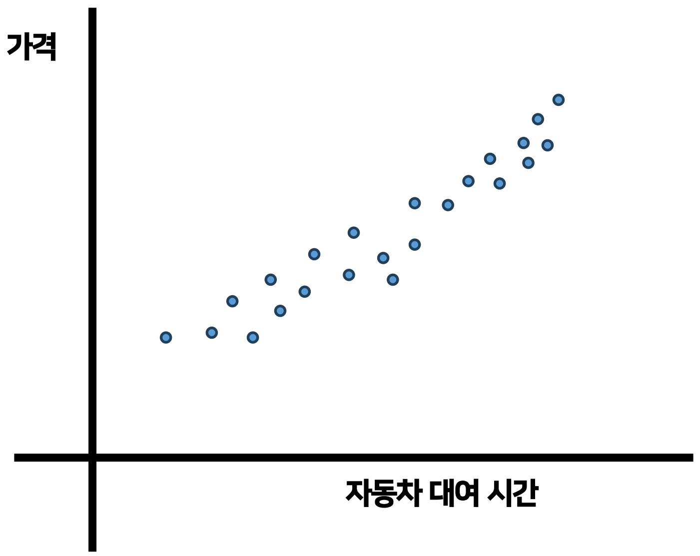
회귀분석(Regression Analysis)



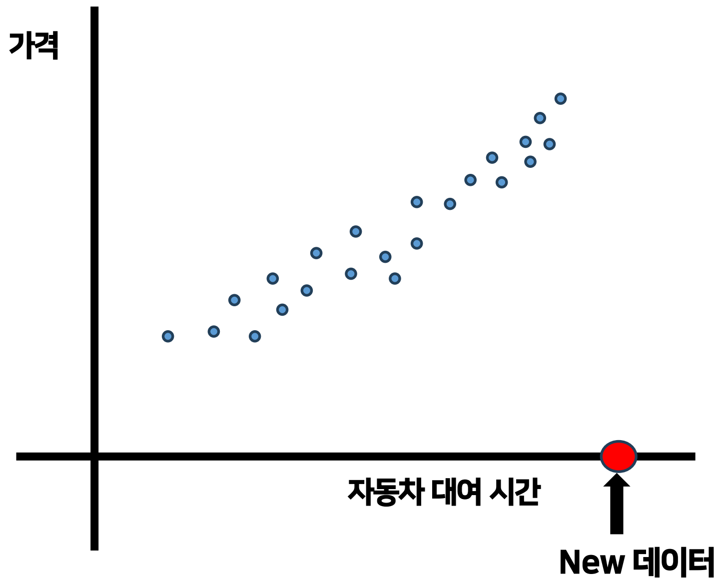
회귀분석(Regression Analysis) vs 통계적 예측(Statistical prediction)

Q) 독립변수의 값이 증가 또는 감소할 수록 종속변수의 값도 증가 또는 감소할 수 있음

추론통계 : 집단의 현상황을 파악하고, 추론해 결과를 도출 하는 방법
→ 가격과 자동차 대여 시간의 서로 상관관계가 존재함
→ 자동차 대여 시간은 가격에 영향을 줌

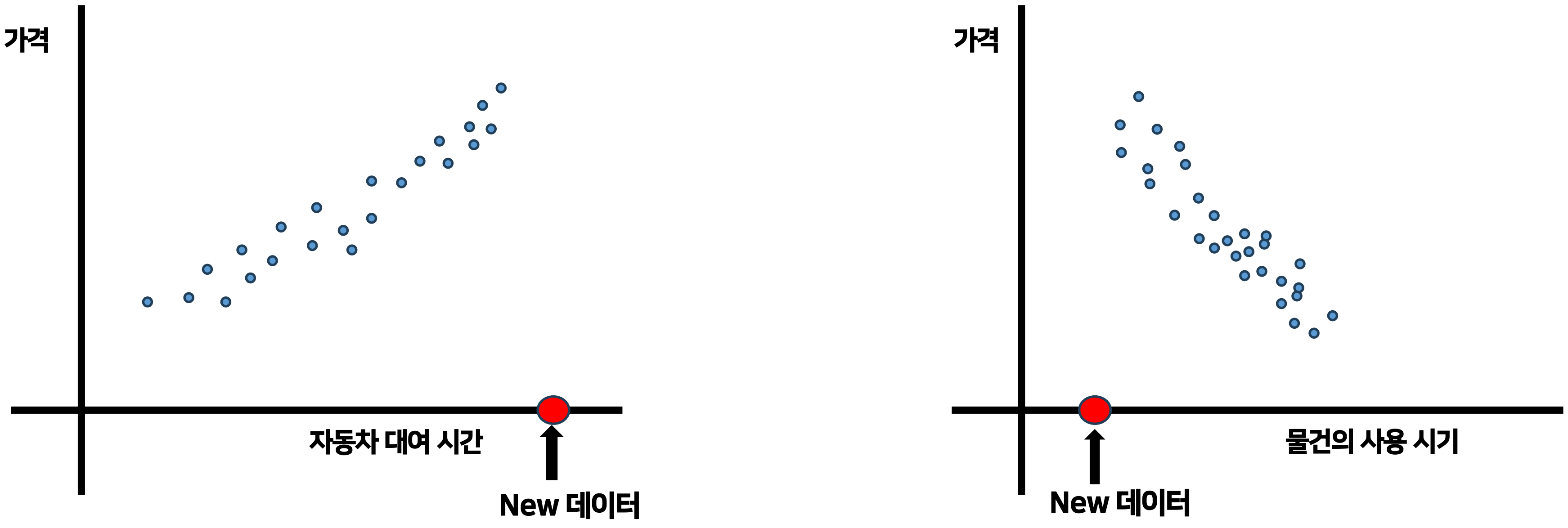


기계학습 : 집단의 현상황을 학습하고, 새로운 상황이 왔을 때, 예측하는 방법
→ 대여시간을 입력할 경우 이에 대한 가격을 예측함
→ 대여시간이 2시간이므로 가격은 20,000원임



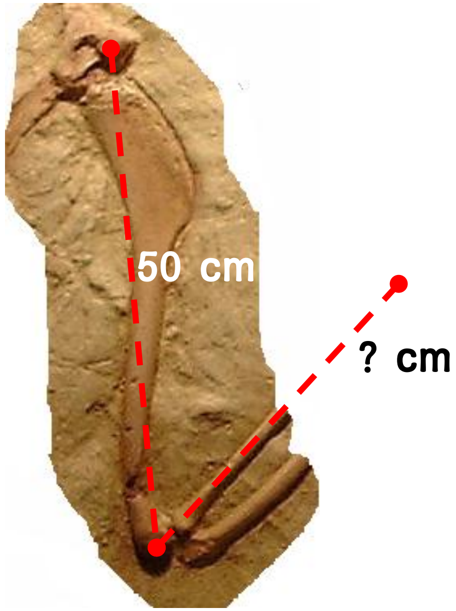
회귀분석(Regression Analysis)

Q) 독립변수의 값이 증가 또는 감소할 수록 종속변수의 값도 증가 또는 감소할 수 있음



회귀분석(Regression Analysis)

Q) 완전하지 않은 화석의 대퇴부의 길이가 50 cm일 때
상박부의 길이는 얼마나 될까? → 기존에 학습했던 데이터를 기반으로 예측가능

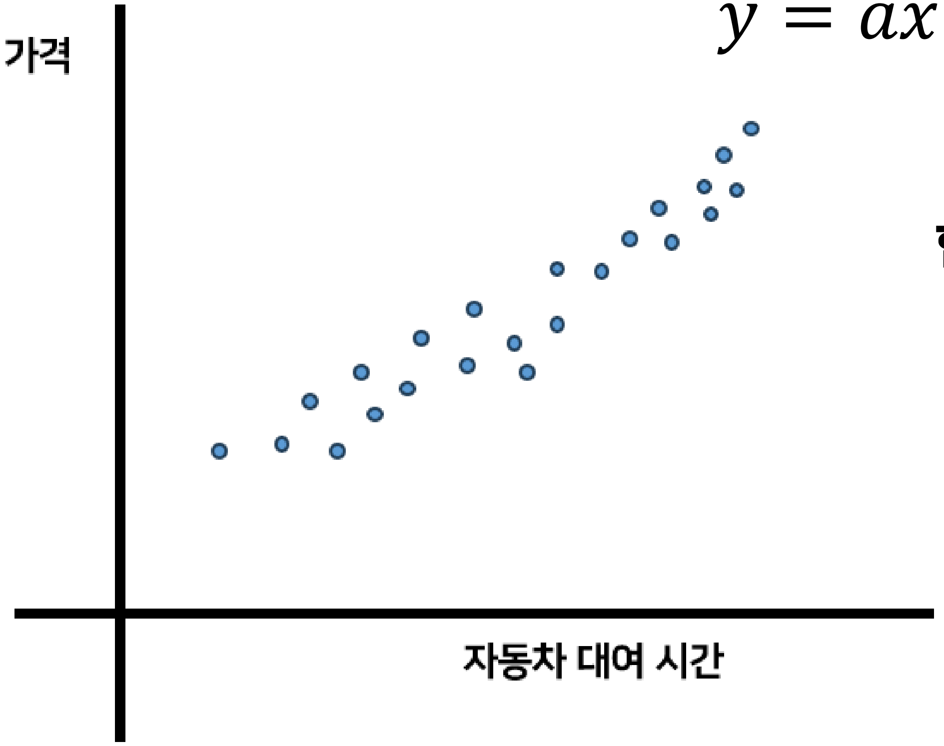


- 대퇴부와 상박부는 어떤 관계가 있을까?
  → 기존의 학습 데이터로 특징을 파악함
- 대퇴부와 상박부가 어떤 관계가 있다면 관계를 수학적 함수로 나타낼 수 있는가?
  → 함수를 만들어 학습하고, 결과를 도출함

02 회귀(Regression)

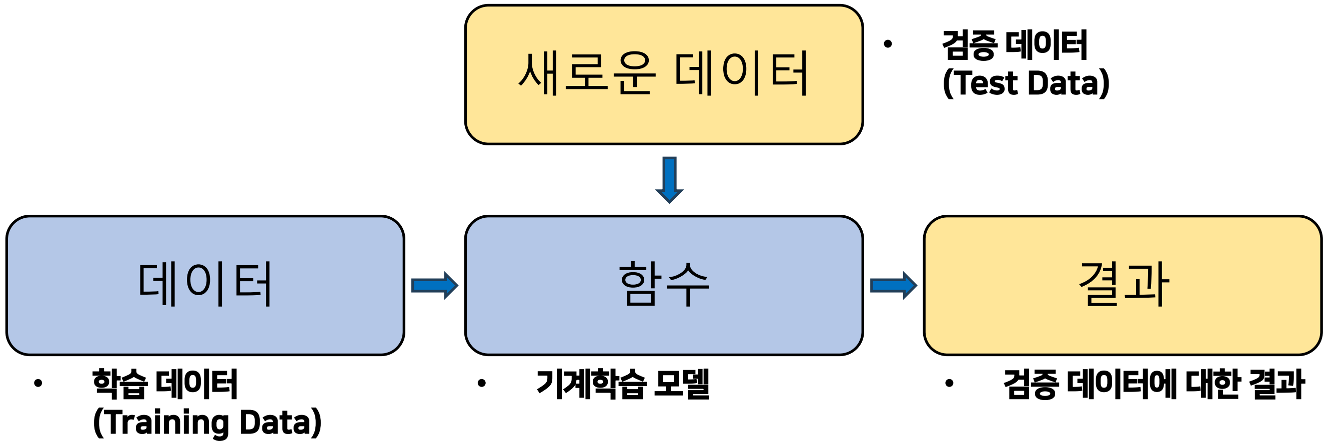
회귀분석(Regression Analysis)

회귀분석의 구조



$$y = ax + b$$

함수(Function)==모델(Model)



회귀분석(Regression Analysis)

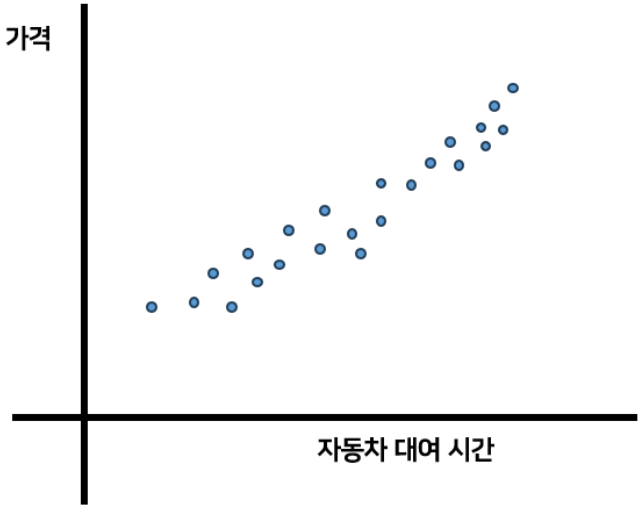
회귀분석의 구조

- 자연현상이나 사회현상에 관련된 하나의 종속변수(또는 반응변수)와 하나 이상의 독립변수(또는 설명변수)의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정함

- 변수들의 자료로부터 가정된 모형의 미지의 회귀계수를 추정하여 현상을 설명하고 예측하는 통계적 분석 방법

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

종속변수 상수 회귀계수 독립변수



회귀분석(Regression Analysis)

회귀분석의 구조 → 회귀계수

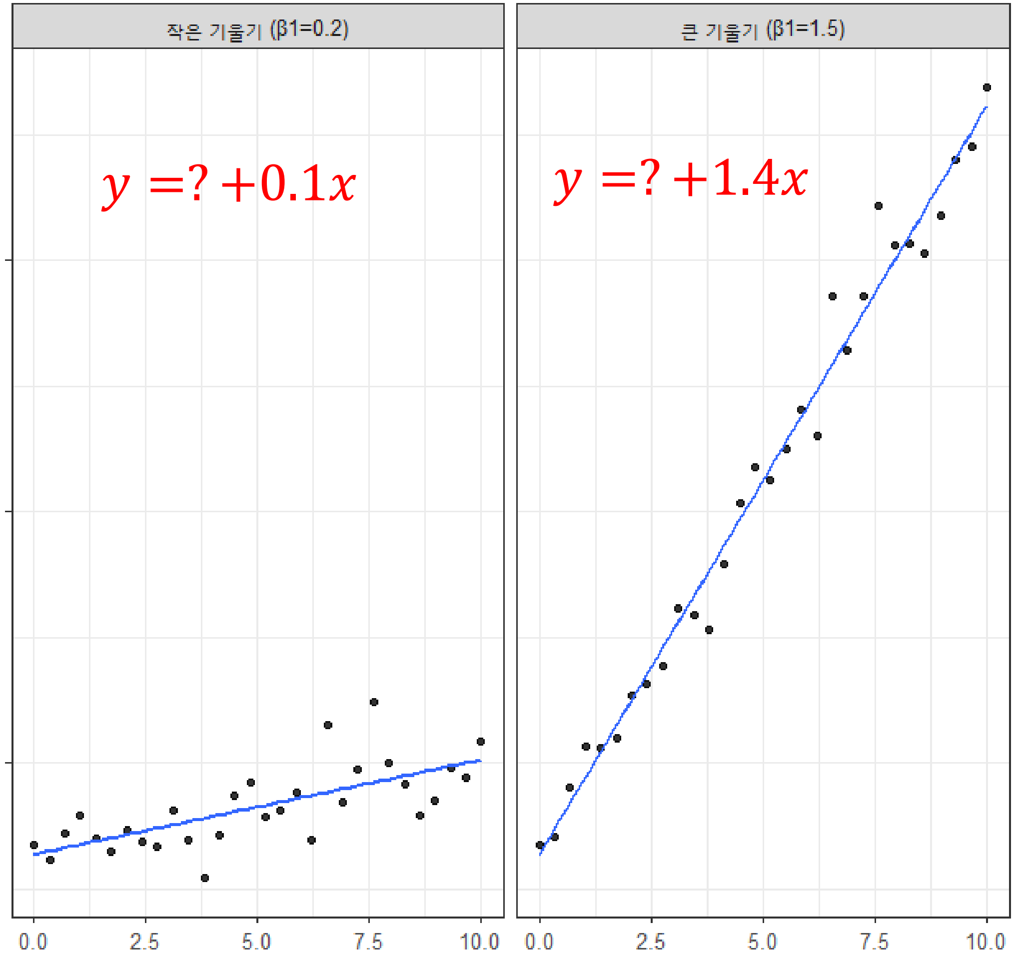
추정된 회귀계수(Estimator of regression coefficient)

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- 독립 변수(x)가 1단위 증가할 때 종속 변수(y)의 평균 변화를 알려줌
- β1이 양수이면 x가 증가함에 따라 y도 증가하는 경향을 가짐
- β1이 음수이면 x가 증가함에 따라 y는 감소하는 경향을 가짐

회귀분석(Regression Analysis)

회귀분석의 구조 → 회귀계수(β_1)



평균

X	Y(작은기울기)
0.0	2.9
0.3	3.5
0.7	2.3
1.0	2.8
1.4	3.2
1.7	4.6
2.1	4.3
2.4	3.4
2.8	4.0
3.1	3.7
3.4	3.1
3.8	3.3
4.1	3.2
2.1	**3.4**

X	Y(큰기울기)
0.0	2.9
0.3	4.0
0.7	3.2
1.0	4.2
1.4	5.0
1.7	6.9
2.1	6.9
2.4	6.5
2.8	7.6
3.1	7.8
3.4	7.5
3.8	8.2
4.1	8.6
2.1	**6.1**

회귀분석(Regression Analysis)

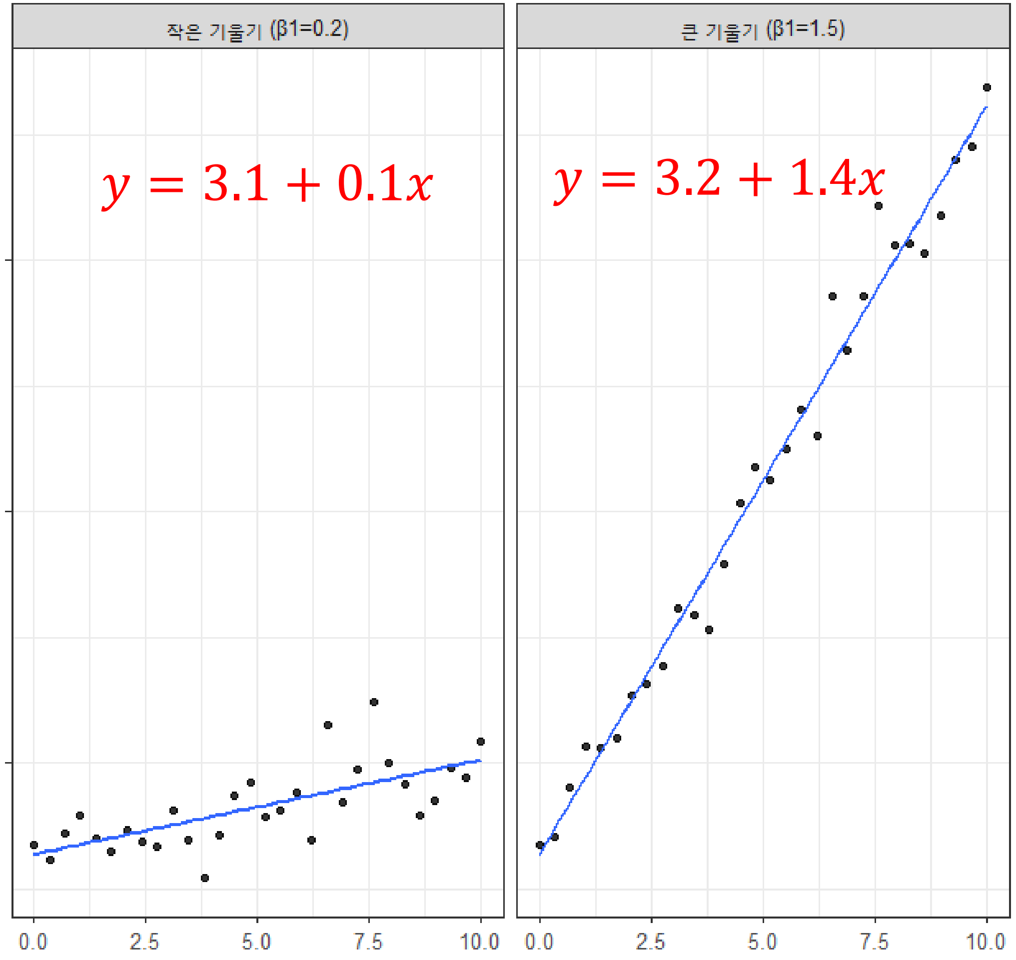
회귀분석의 구조 → 상수

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- 절편(β0)은 모든 독립변수가 없거나 0인 경우 종속변수의 기준 수준을 제공함
- 가장 적합한 선을 제공하는 절편과 기울기의 최소 제곱 추정치를 계산함
- 계수(β1,β2,⋯)는 각 독립변수가 종속변수에 얼마나 많은 방향으로 영향을 미치는지 보여줌

회귀분석(Regression Analysis)

회귀분석의 구조 → 회귀계수 (β_0)



평균

X	Y(작은기울기)
0.0	2.9
0.3	3.5
0.7	2.3
1.0	2.8
1.4	3.2
1.7	4.6
2.1	4.3
2.4	3.4
2.8	4.0
3.1	3.7
3.4	3.1
3.8	3.3
4.1	3.2
2.1	**3.4**

X	Y(큰기울기)
0.0	2.9
0.3	4.0
0.7	3.2
1.0	4.2
1.4	5.0
1.7	6.9
2.1	6.9
2.4	6.5
2.8	7.6
3.1	7.8
3.4	7.5
3.8	8.2
4.1	8.6
2.1	**6.1**

회귀분석(Regression Analysis)

회귀분석의 구조 → 오차항(관찰 불가능)

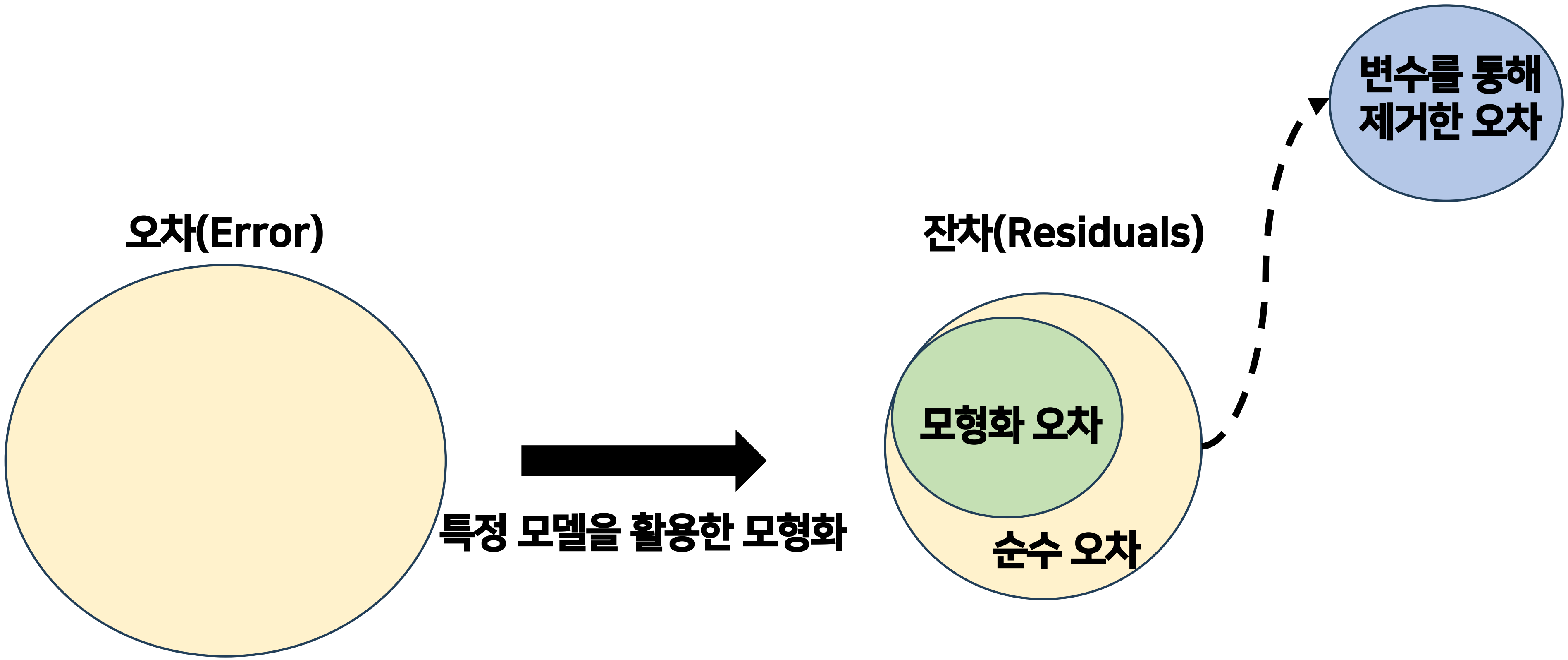
$$y = \beta_0 + \beta_1 x + \varepsilon$$

가정 : $\varepsilon_i \sim iidN(0, \quad \sigma^2)$

- 오차(Error) : 데이터가 실제로 만들어질 때 섞여 들어가는 본질적인 잡음
  → 보이지 않고, 생성 규칙을 모르기 때문에 직접 측정할 수 없음 → 모델에 포함되면
- 모형화 오차 : 오차가 모델에 포함되어 오차에 들어있던 체계적 요인이 빠져나와 모형화 오차로 변화함
- 잔차(Residuals) : 관찰값과 모델의 예측값 사이의 차이로, 우리가 실제로 계산해서 볼 수 있는 값
- 잔차 = 오차(모형화 오차 뒤 남은 순수한 설명 불가 부분) + 모형화 오차(모델의 미스매치)

회귀분석(Regression Analysis)

회귀분석의 구조 → 오차항(관찰 불가능)

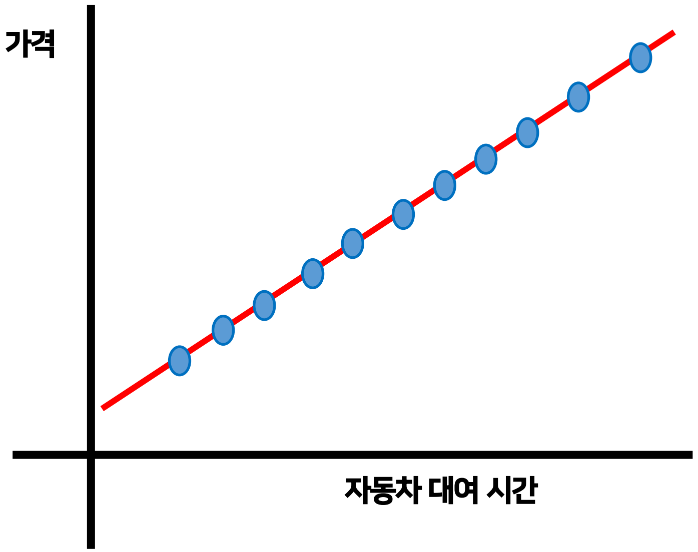


회귀분석(Regression Analysis)

회귀분석의 구조 → 오차항(관찰 불가능)

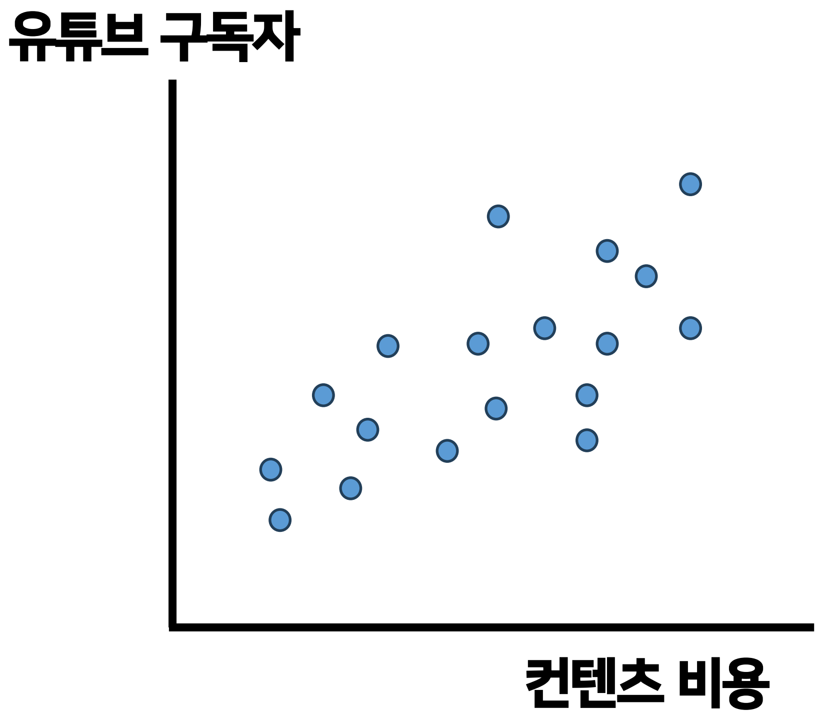
$$y = \beta_0 + \beta_1 x + \varepsilon$$

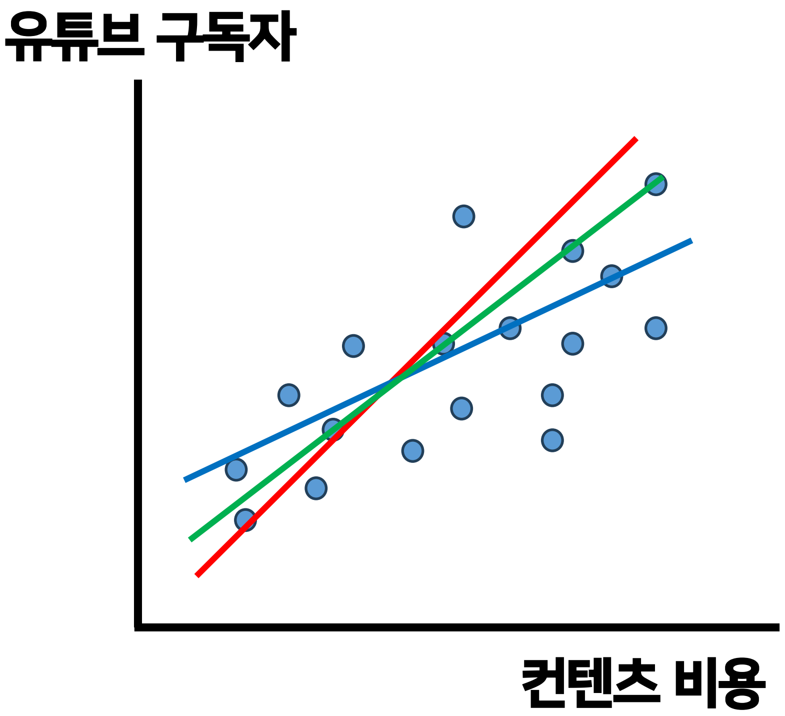
$$y = \beta_0 + \beta_1 x$$



회귀분석(Regression Analysis)

최소제곱법(Ordinary Least Squares) : SSE(Sum of Squared Residuals/Error)를 최소로 하는 직선을 구하는 것

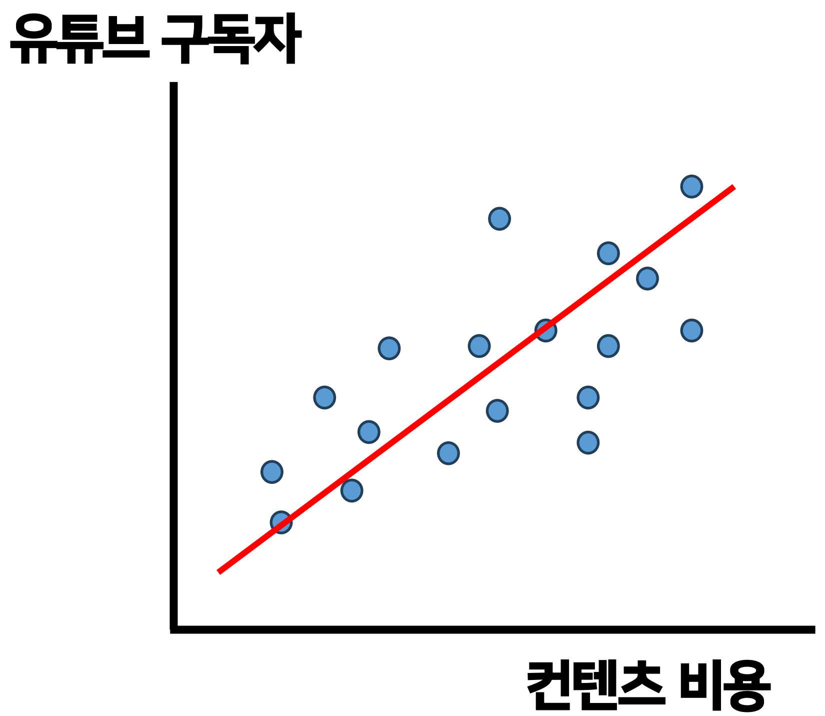


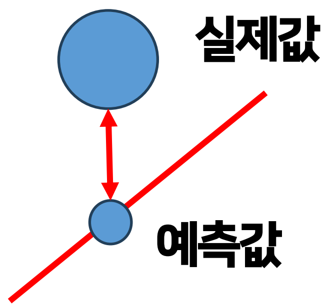


회귀분석(Regression Analysis)

회귀분석의 구조 → 상수 및 회귀계수

오차 제곱합을 최소로 하는 β0 및 β1 을 찾는 것
SSE(Sum of Squared Residuals/Error) : 적합치에서 반응 값의 전체 잔차(Residuals)를 측정함





$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

실제값 (Y_i)	예측값 $(\widehat{Y_i})$	SSE $(Y_i - \widehat{Y}_i)^2$
4	6	4
10	5	25
5	7	4
12	5	49
...	...	...
		SSE=500

회귀분석의 구조 → 상수 및 회귀계수

오차 제곱합을 최소로 하는 β0 및 β1 을 찾는 것

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0$$

실제값(X_i)	$X_i - 10$	실제값(Y_i)	$Y_i - 30$
4	-6	4	-26
10	0	25	-5
5	-5	4	-26
12	2	49	19
...	...	...	...
평균 : 10		평균=30	

회귀분석(Regression Analysis)

회귀분석의 구조 → 오차항(관찰 불가능)

오차항의 특징

- 완전성 : 모델이 모든 것을 포착하지 못함
- 복잡성 : 데이터 제한 또는 알 수 없는 요인이 결과에 영향 미치는 모든 변수를 모델에 포함할 수는 없음
- 유연성: 설명할 수 없는 구성 요소가 있고, 독립 변수와 종속 변수 간의 체계적인 관계를 포착하는 데 집중
- 통찰력 : 모델에서 누락된 부분에 대한 통찰력을 얻을 수 있으며, 개선 및 탐색할 추가 변수를 제안할 수 있음

오차항의 예

- 위치 : 더 바람직한 지역에 있는 집은 더 저렴한 집과 크기가 같더라도 더 많은 비용이 들 수 있음
- 조건 : 잘 관리된 주택은 상당한 수리가 필요한 비슷한 크기의 주택보다 더 높은 가격에 판매될 수 있음
- 시장 동향 : 주택 시장의 변동은 주택의 특성과 관계없이 가격에 영향을 미칠 수 있음

회귀분석(Regression Analysis)

회귀분석의 구조 → 잔차(관찰 가능)

- 오차항(ϵ) : 모델에 없는 요인(예: 예상치 못한 도로 폐쇄)으로 인해 발생한 차이를 나타냄
  → 특정 날짜에 이러한 모든 요소를 미리 알거나 정확한 영향을 정량화할 수 없음
- 잔차(e) : 예측이 실제로 관찰한 것과 얼마나 다른지 나타내고, 모델이 경험한 것과 다른 시간을 일관되게 예측하는 경우 잔차를 조사하면 모델을 이해하고 개선하는 데 도움이 될 수 있음
  - 어느 월요일, 오전 8시에 출발했는데 모델이 과거 데이터를 기반으로 30분 운전을 예측했지만 실제로는 35분이 소요됨
  - 이 이동의 잔여 시간 ϵ 은 +5분(실제 35분 - 예측 30분), 화요일에는 조건이 조금 다르지만 동시에 출발하면 28분이 걸려 잔여시간이 -2분(실제 28분~예상 30분)이 됨

회귀분석(Regression Analysis)

회귀분석의 구조 → 잔차(관찰 가능)

어느 월요일, 오전 8시에 출발했는데 모델이 과거 데이터를 기반으로 30분 운전을 예측했습니다. 그 [illegible]
화요일에는 조건이 조금 다르지만 동시에 출발하면 28분이 걸려 잔여시간이 -2분(실제 28분~예상 30분)이 됩니다.

회귀분석은 독립변수가 종속변수에 미치는 영향을 수학적 식으로 모델링해 예측·분석하는 통계 기법

예시된 핵심 사항

→ 하나의 선을 통해 최소의 잔차를 가지는 모델

오차항(ϵ): 모델에 없는 요인(예: 예상치 못한 도로 폐쇄)으로 인해 발생한 차이를 나타냅니다. 특정 날짜에 이러한 모든 요소를 미리 알거나 정확한 영향을 정량화할 수 없기 때문에 이는 이론적

잔차(e): 예측이 실제로 관찰한 것과 얼마나 다른지 나타내고, 모델이 경험한 것과 다른 시간을 일관되게 예측하는 경우 잔차를 조사하면 모델을 이해하고 개선하는 데 도움이 될 수 있음

회귀분석(Regression Analysis)

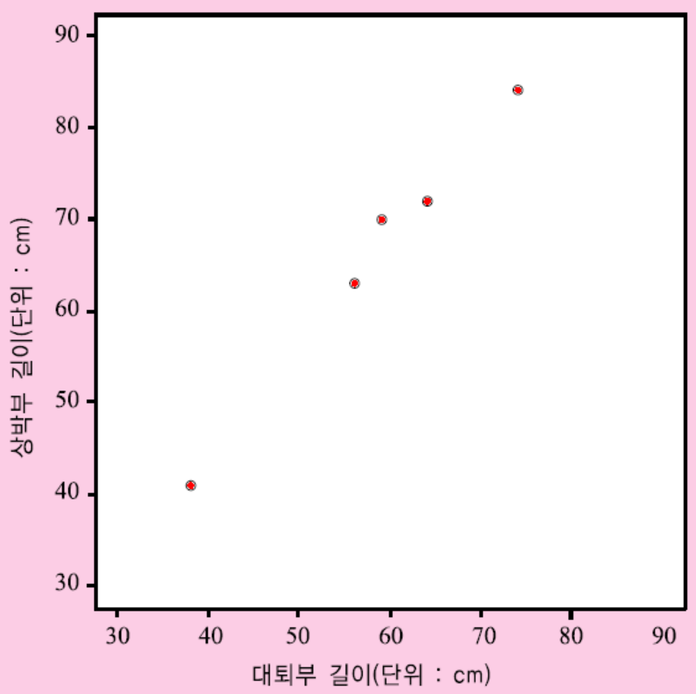
회귀분석은 많은 분야에서 데이터 분석에 널리 사용되는 통계적 방법 중 하나

- 변수 간의 관계 파악 : 회귀분석은 종속변수와 독립변수 사이의 관계를 파악하는데 유용하고, 이를 통해 변수 간의 인과 관계를 파악하거나, 변수 간의 연관성을 파악할 수 있음
- 예측 모델 구축 : 회귀분석은 종속변수와 독립변수 간의 관계를 이용하여 예측 모델을 구축하는 데 사용되고, 이를 통해 미래의 값을 예측하거나, 어떤 조건에서의 종속변수의 값 변화를 예측할 수 있음
- 변수의 영향력 파악 : 회귀분석은 독립변수가 종속변수에 미치는 영향력을 파악하는데 사용되고, 이를 통해 어떤 독립변수가 종속변수에 가장 큰 영향력을 미치는지 파악할 수 있음
- 이상치 탐지 : 회귀분석은 이상치를 탐지하는데도 사용되고, 이상치는 예측 모델의 정확도를 낮추거나 분석 결과를 왜곡시킬 수 있으므로, 회귀분석을 통해 이상치를 탐지하고 제거할 수 있음

회귀분석(Regression Analysis) - 단순회귀분석

시조새 5개의 화석에 대하여 대퇴부(다리의 뼈)와 상박부(팔 윗부분의 뼈)의 길이(단위:cm)를 측정한 자료이다.

대퇴부	38	56	59	64	74
상박부	41	63	70	72	84



- **대퇴부와 상박부는 어떤 관계가 있을까?**

직선적 관계

- **대퇴부와 상박부가 어떤 관계가 있다면 관계를 수학적 함수로 나타낼 수 있는가?**

$$상박부 = \beta_0 + \beta_1(대퇴부)$$

회귀분석(Regression Analysis) - 단순회귀분석

수학적 함수관계 :

상박부 길이 $= \beta_0 + \beta_1 \times$ 대퇴부 길이

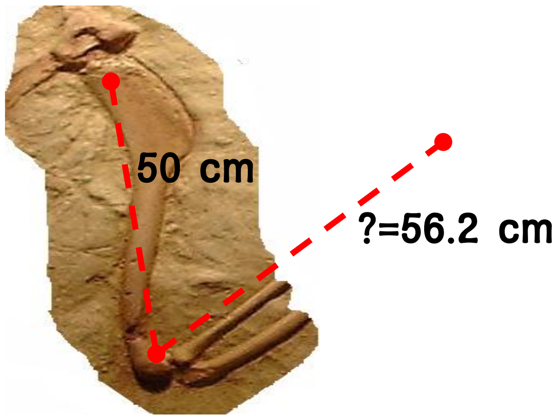
상박부 길이= **-3.66** + **1.197** × 대퇴부 길이

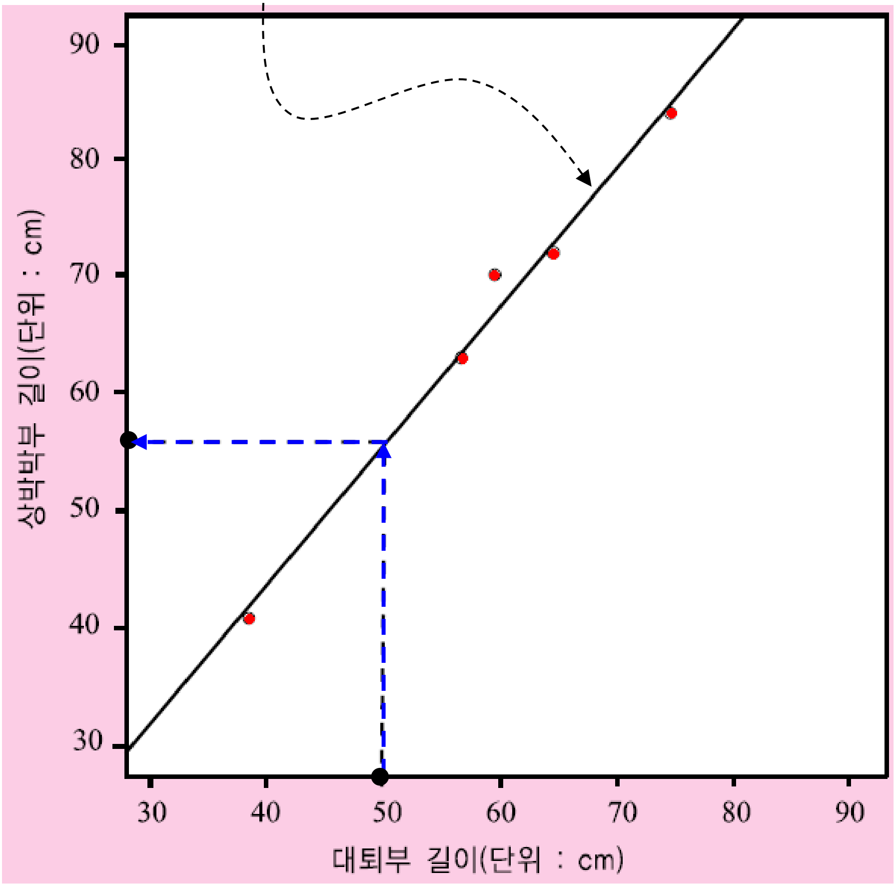
통계적 예측 :

예측 상박부의 길이

= -3.66+(1.197)(50)

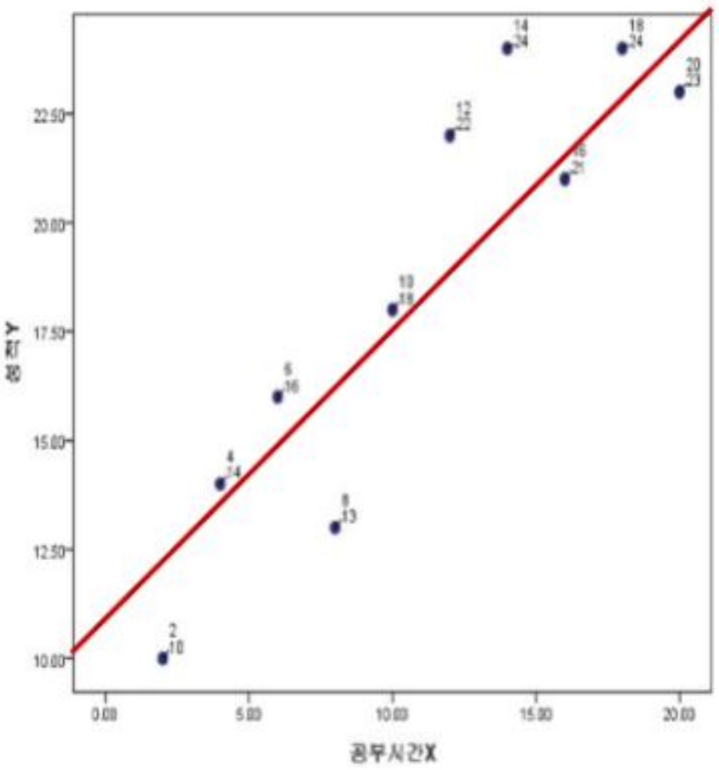
= 56.2cm





회귀분석(Regression Analysis) - 단순회귀분석

Q) 공부시간과 성적의 관계에 대한 관계에 대해서 알아보자.

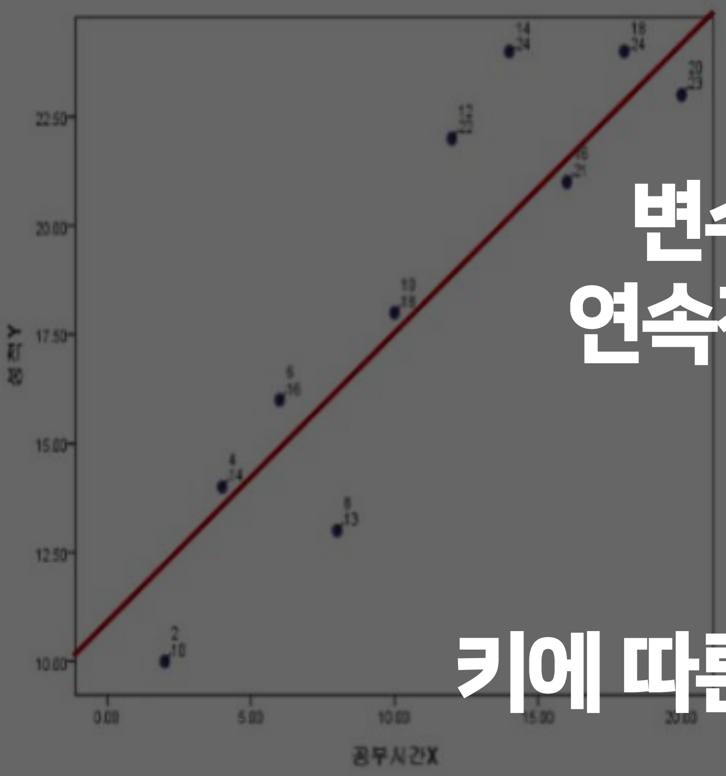


- 공부의 시간과 성적은 어떤 관계가 있을까?

- 공부와 성적의 관계에 대해 수치적으로 정의를 내릴수가 있을까?

회귀분석(Regression Analysis) - 단순회귀분석

Q) 공부시간과 성적의 관계에 대한 관계에 대해서 알아보자.



- 공부의 시간과 성적은 어떤 관계가 있을까?

변수들 간의 상관관계를 찾는 것
연속적인 데이터로부터 결과를 예측

- 공부와 성적의 관계에 대해 수치적으로 정의를 내릴수가 있을까?

키에 따른 몸무게, 공부시간에 따른 점수 …

회귀분석(Regression Analysis) - 단순회귀분석

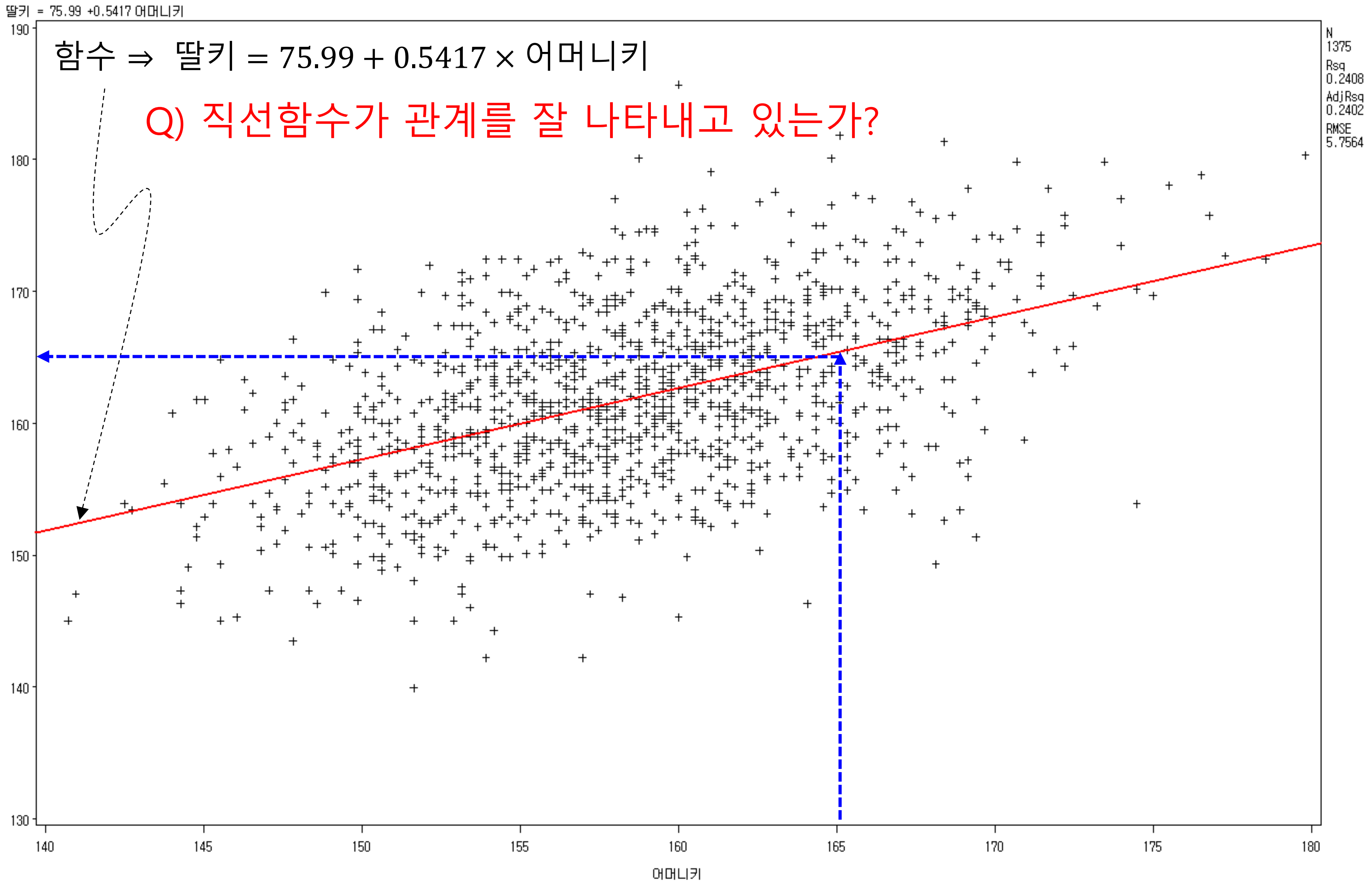
어머니(65미만)와 딸(18세 이상)의 신장(cm)

Q) 어머니 키와 딸 키에는 어떤 관계가 있는가?

Q) 관계가 있다면 어떤 함수로 관계를 나타낼 수 있는가?

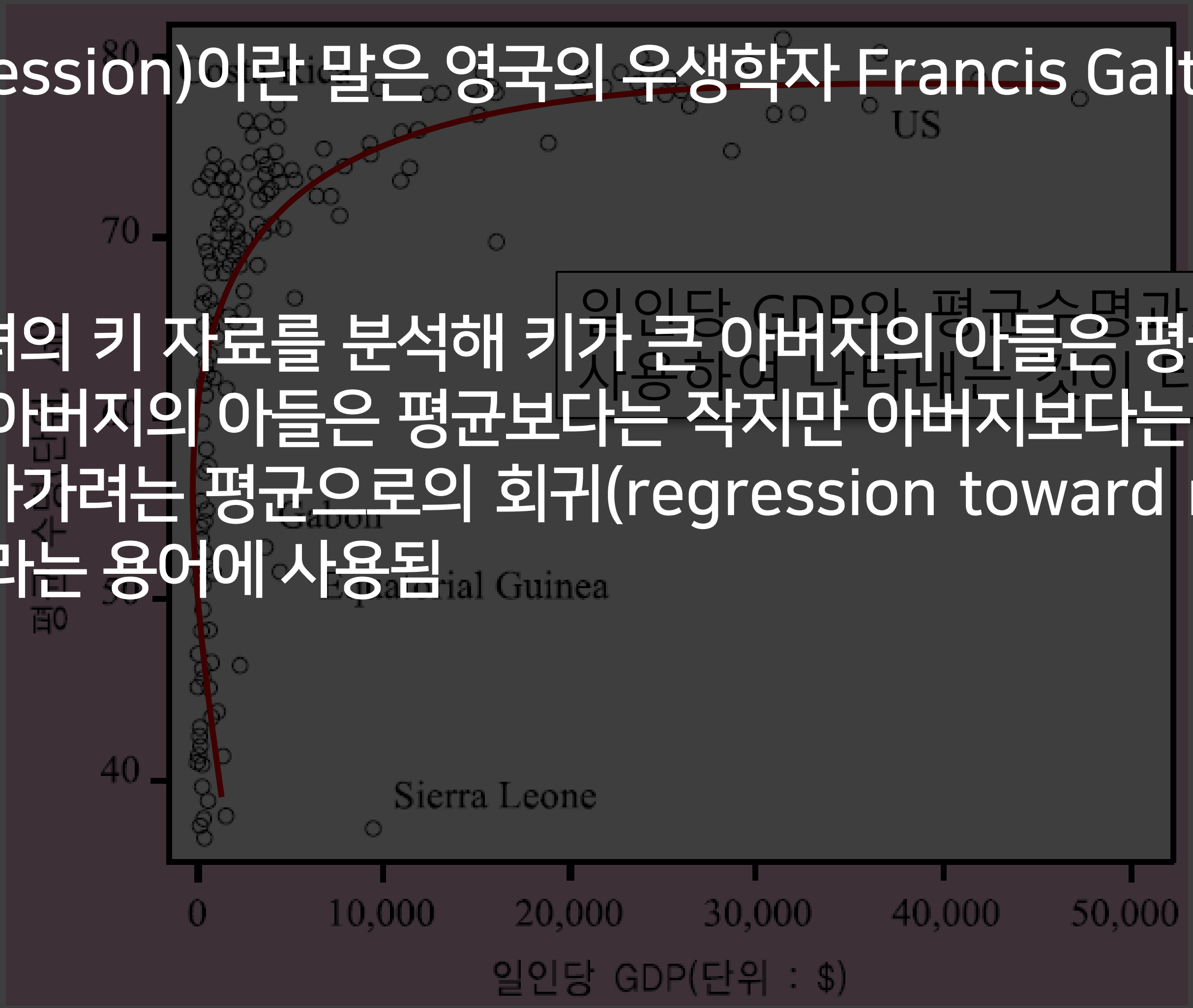
회귀분석(Regression Analysis) - 단순회귀분석

아버지(65세 미만)와 아들(18세 이상)의 신장(cm)



회귀분석(Regression Analysis) - 단순회귀분석

평균 수명과 일인당 GDP를 비교



회귀(regression)이란 말은 영국의 우생학자 Francis Galton이 처음 사용함

부모와 자녀의 키 자료를 분석해 키가 큰 아버지의 아들은 평균보다는 크지만 아버지보다는 작고, 키가 작은 아버지의 아들은 평균보다는 작지만 아버지보다는 큰 경향을 보이는 현상을 결국 평균 키로 되돌아가려는 평균으로의 회귀(regression toward mediocrity)라고 불렀고 이는 오늘날의 회귀 라는 용어에 사용됨

회귀분석(Regression Analysis)

어떤 관련이 있는가?

고장난 컴퓨터의 부품의 수와 수리시간과의 관계를 알기 위해

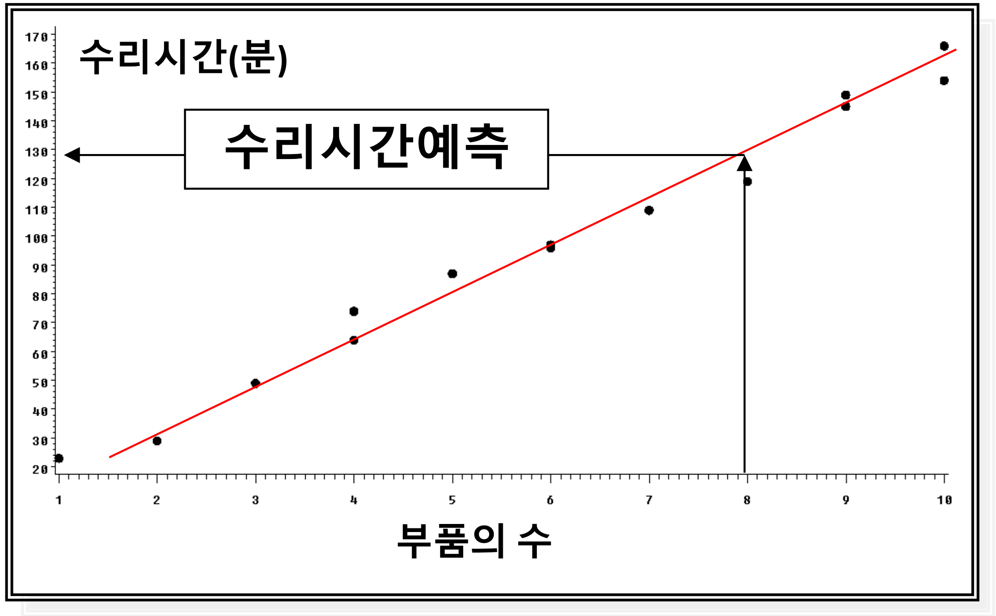
조사한 자료의 산점도는 다음과 같음

어떠한 함수가 고장난 부품의 수와 수리시간의 관계를 잘 나타낼 수 있는가?

직선함수

수리시간=ax(부품의 수)+b

추정(estimation)



02 회귀(Regression)

회귀분석(Regression Analysis)

종 류	설 명	수학적 모 형
단순회귀 Simlpe Linear Reg	독립변수(x)가 하나이고 종속변수(y)와의 관계가 직선이다.	$y = \beta_0 + \beta_1 x + \varepsilon$
다중회귀 Multiple Linear Reg	독립변수가 2개 이상 ($x_1,\ldots,x_k$)이고 종속변수 (y)와의 관계가 1차선형 함수	$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$
곡선회귀 Curvilinear Linear Reg	독립변수(x)가 하나이고 종속변수(y)와의관계가 2차곡선 이상	K차 곡선모형 $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$
다항회귀 Polynomial Linear Reg	중회귀와 곡선회귀를 합친 것	독립변수 2개인 2차 곡선모형 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$
비선형회귀 Nonlinear Reg	회귀계수들이 선형으로 표현되지 않는경우	$y = \beta_0 \beta_1 x + \varepsilon$ $y = \beta_0 + \beta_1^2 x + \varepsilon$

회귀분석(Regression Analysis)

회귀분석의 프로세스

(1) 독립변수와 종속변수에 대한 자료의 산점도를 그림

(2) 산점도로 부터 적절한 회귀모형을 선택

(3) 선택된 회귀모형을 자료를 이용하여 회귀계수를 추정

(4) 회귀모형의 적용이 옳은가를 판단

실습

```
#데이터 불러오기
train_data <- read.csv("simple_train_data.csv")

#산점도
library(ggplot2)

ggplot(train_data, aes(x = thigh, y = upper_arm)) +
  geom_point(alpha = 0.7) +
  labs(title = "대퇴부 vs 상박부 산점도",
       x = "대퇴부 둘레 (thigh, cm)",
       y = "상박부 둘레 (upper_arm, cm)") +
  theme_minimal()

#회귀분석 모델
model <- lm(upper_arm ~ thigh, data = train_data)
summary(model)

#잔차
residuals(model)
```

실습

```
Call:
lm(formula = upper_arm ~ thigh, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4069 -0.9916 -0.0375  0.9879  4.6095 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.784231   0.324304   17.84   <2e-16 ***
thigh       0.388886   0.005374   72.36   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.489 on 998 degrees of freedom
Multiple R-squared:  0.8399,	Adjusted R-squared:  0.8398 
F-statistic:  5237 on 1 and 998 DF,  p-value: < 2.2e-16
```

귀무가설 : 상수가 0이다.

귀무가설 : x와y는 선형 관련이 없다.

실습

```
Residual standard error: 1.489 on 998 degrees of freedom
Multiple R-squared:  0.8399,    Adjusted R-squared:  0.8398
F-statistic:  5237 on 1 and 998 DF,  p-value: < 2.2e-16
```

귀무가설 : 독립변수는 종속변수를 유의미하게 설명하지 않는다.

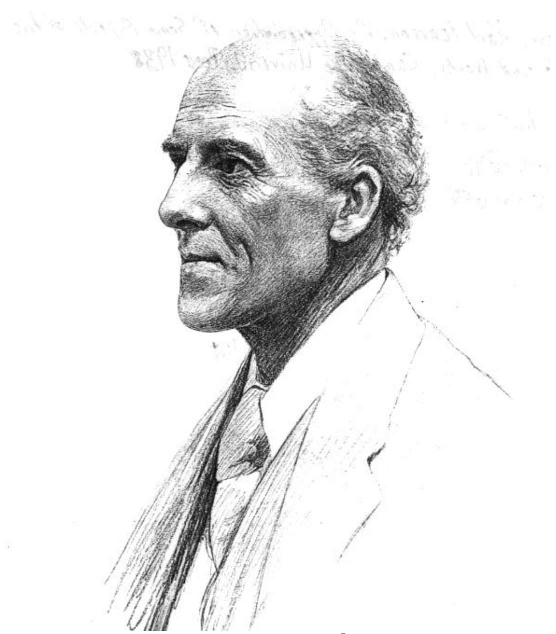
- 0에 가까움: 모형이 종속변수를 거의 설명 못함
- 1에 가까움: 모형이 종속변수를 거의 완벽히 설명
- 0.3 ~ 0.5: 사회과학/행동과학 등에서 “쓸 만하다”는 수준
- 0.7 이상: 설명력이 강하다고 평가
- 0.9 이상: 설명력이 매우 강하다고 평가

실습

```r
#데이터 불러오기
train_data <- read.csv("simple_train_2.csv")

#산점도
library(ggplot2)

ggplot(train_data, aes(x = thigh, y = upper_arm)) +
  geom_point(alpha = 0.7) +
  labs(title = "대퇴부 vs 상박부 산점도",
       x = "대퇴부 둘레 (thigh, cm)",
       y = "상박부 둘레 (upper_arm, cm)") +
  theme_minimal()

#회귀분석 모델
model <- lm(upper_arm ~ thigh, data = train_data)
summary(model)

#잔차
residuals(model)
```

상관관계(Correlation)

두 변수(양적변수) x와 y 사이의 직선적인 상관관계의 정도를 표본으로부터 수량적으로 표현한 값을 표본상관계수(sample coefficient of correlation) 또는 피어슨 상관계수라 한다.



Pearson, Karl(1857-1936)

상관관계(Correlation)

양적변수 x와 y에 대한 자료 : $(x_1, \quad y_1), \quad (x_2, \quad y_2), \quad \cdots, \quad (x_n, \quad y_n)$

$\rho(X, Y)$

$$\boxed{r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

상관관계(Correlation)

상관관계를 알기 위해서는

$(x_i - \bar{x})$ ← 편차

$(x_i - \bar{x})^2$ ← 편차 제곱

$\sum(x_i - \bar{x})^2$ ← 편차 제곱합

$\frac{\sum(x_i - \bar{x})^2}{n}$ ← 편차 제곱의 평균(분산)

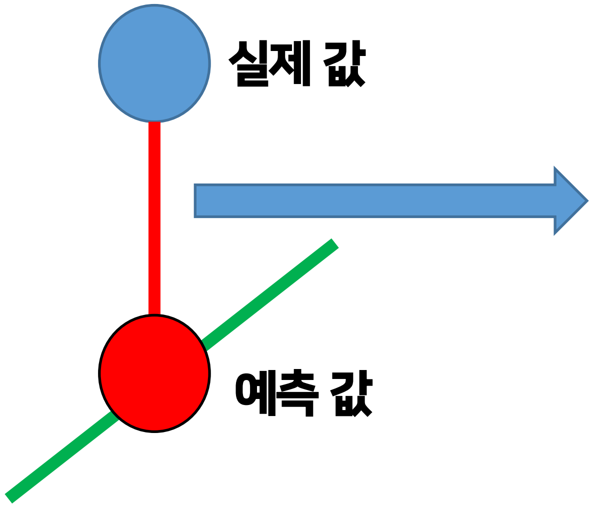
$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$ ← 루트분산(표준편차)

공분산 : 두 변수가 얼마나 같이 움직이는지

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

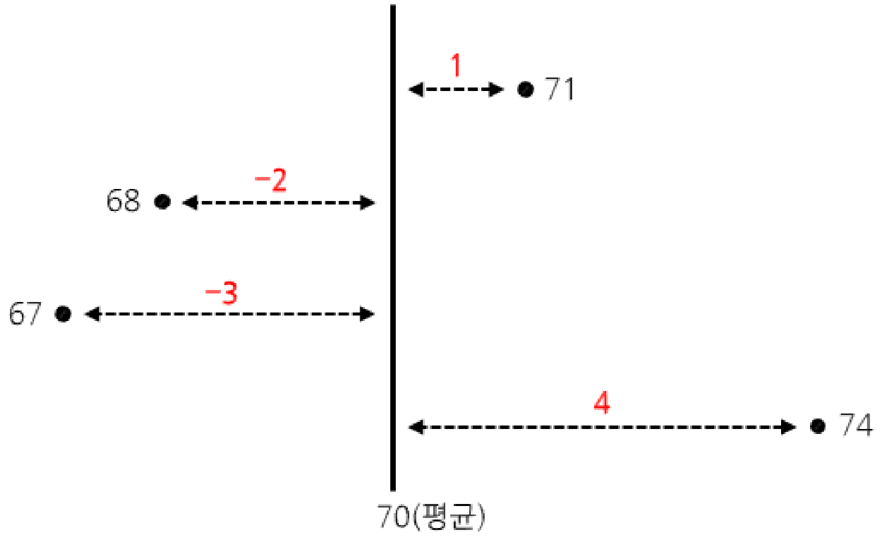
두 변수의 크기(변동성=흩어진 정도)을 고려해 정규화

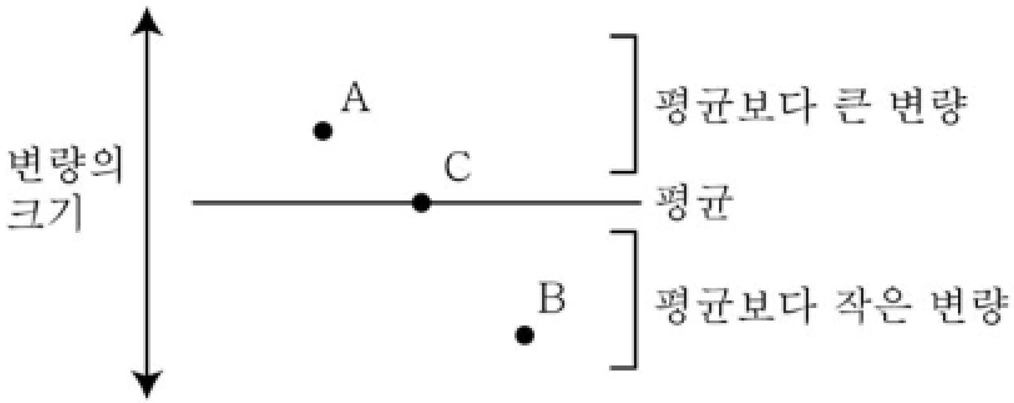
상관관계(Correlation)



편차 제곱 : 실제 값과 예측 값 차이의 제곱의 합을 최소화

편차 : 변량-평균



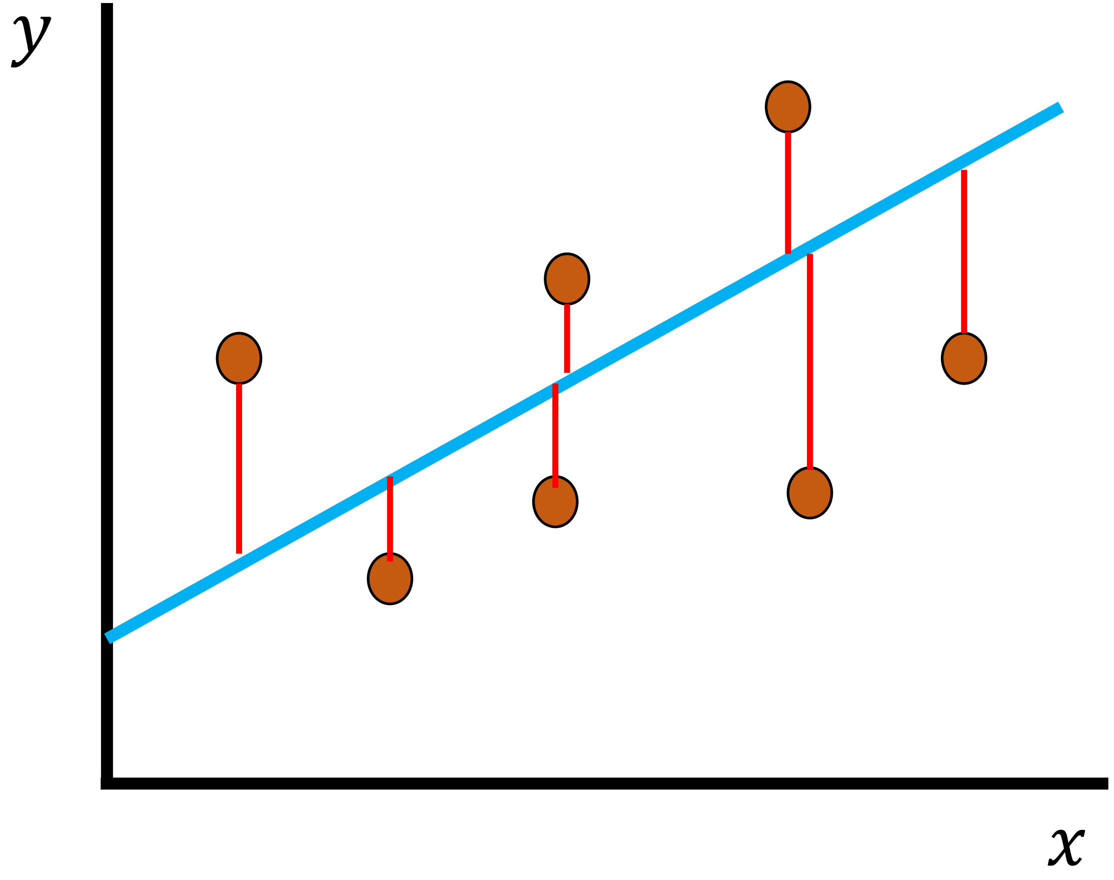


$$\frac{1^2 - 2^2 - 3^2 + 4^2}{4} = 7.5$$

$(x_i - \bar{x})$ ← 편차

상관관계(Correlation)

편차 : 변량-평균



$$y = mx + b$$

m : 기울기(slope, coefficient)
b : y 절편(intercept)

상관관계(Correlation)

분산 : 편차를 각각 제곱한 값들의 평균

흩어진 정도를 나타내는 분산에 있어서 왜 제곱을 할까?

제곱을 하지 않으면 어떤 일이 생길까?

$(x_i - \bar{x})$ ← 편차

$(x_i - \bar{x})^2$ ← 편차 제곱

$\sum(x_i - \bar{x})^2$ ← 편차 제곱합

$\frac{\sum(x_i - \bar{x})^2}{n}$ ← 편차 제곱의 평균(분산)

제곱을 함으로써 작은 값들의 영향력을 강화하고 큰 값들의 영향력을 덜 강화할 수 있음
예를 들어, 값이 10인 데이터와 값이 100인 데이터가 있을 때, 분산을 계산할 때 제곱을 하면 100인 데이터의 영향력이 10인 데이터의 영향력보다 100배 크게 됨

상관관계(Correlation)

분산은 편차가 제곱 된 값이므로 실질적인 치우침에 비해 그 값이 크기 때문에 루트를 씌워 값을 조절함

이를 표준편차

평균 → 편차→ 분산 → 표준편차

$$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}}$$ ← 루트분산(표준편차)

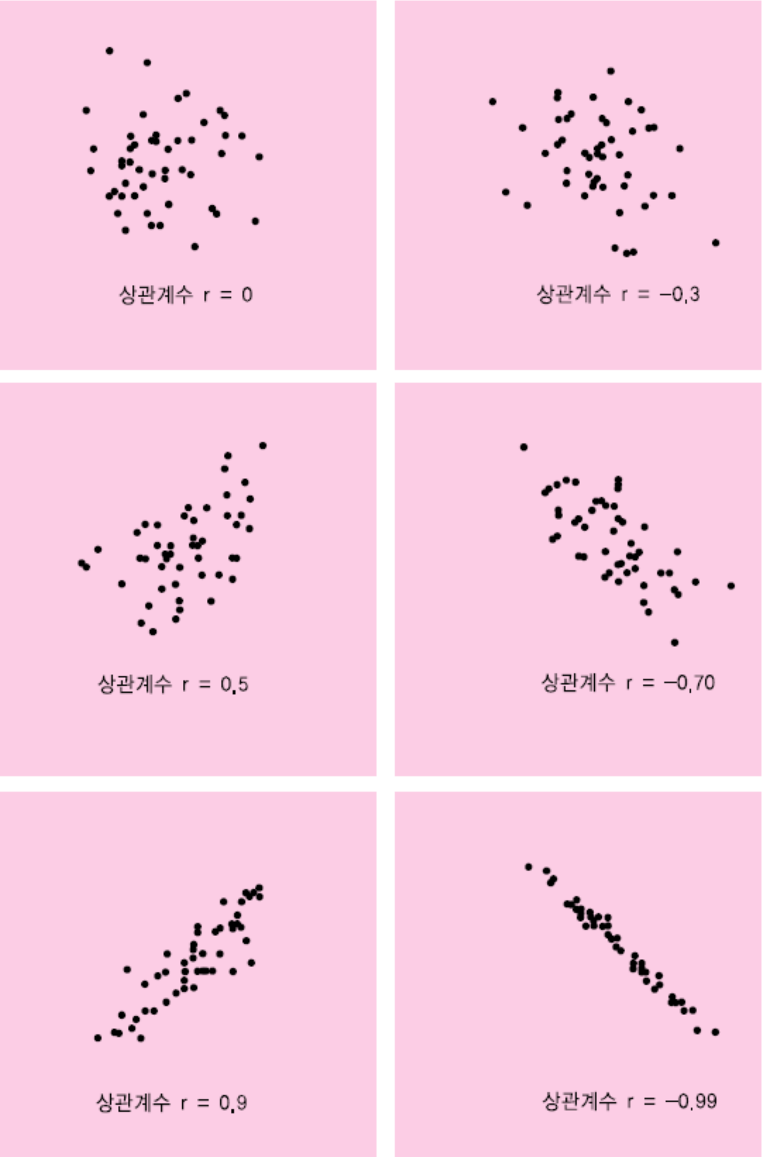
상관관계(Correlation)

- 공분산은 두 변수 X와 Y가 있을 때, X와 Y의 편차를 곱한 값의 평균
- 즉, 공분산은 X와 Y가 동시에 증가하거나 감소하는 경향이 있는지, 아니면 X는 증가하면서 Y는 감소하는 경향이 있는지 등을 나타내는 지표
- 공분산이 양수인 경우, X와 Y가 같은 방향으로 움직이는 경향이 있고, 음수인 경우, X와 Y가 반대 방향으로 움직이는 경향

- 분산은 한 변수의 편차의 제곱의 평균으로 분산은 해당 변수의 변동성을 나타내는 지표
- 분산은 항상 양수이며, 단위는 해당 변수의 단위의 제곱

02 회귀(Regression)

표본상관계수 r의 성질



(1) 표본상관계수의 범위는 $-1 \le r \le 1$이다.

(2) $0 < r \le 1$이면 양의 직선적 상관관계를 갖는다.

(3) $-1 \le r < 0$이면 음의 직선적 상관관계를 갖는다.

(4) $r = 0$이면 직선적 상관관계를 갖지 않는다.

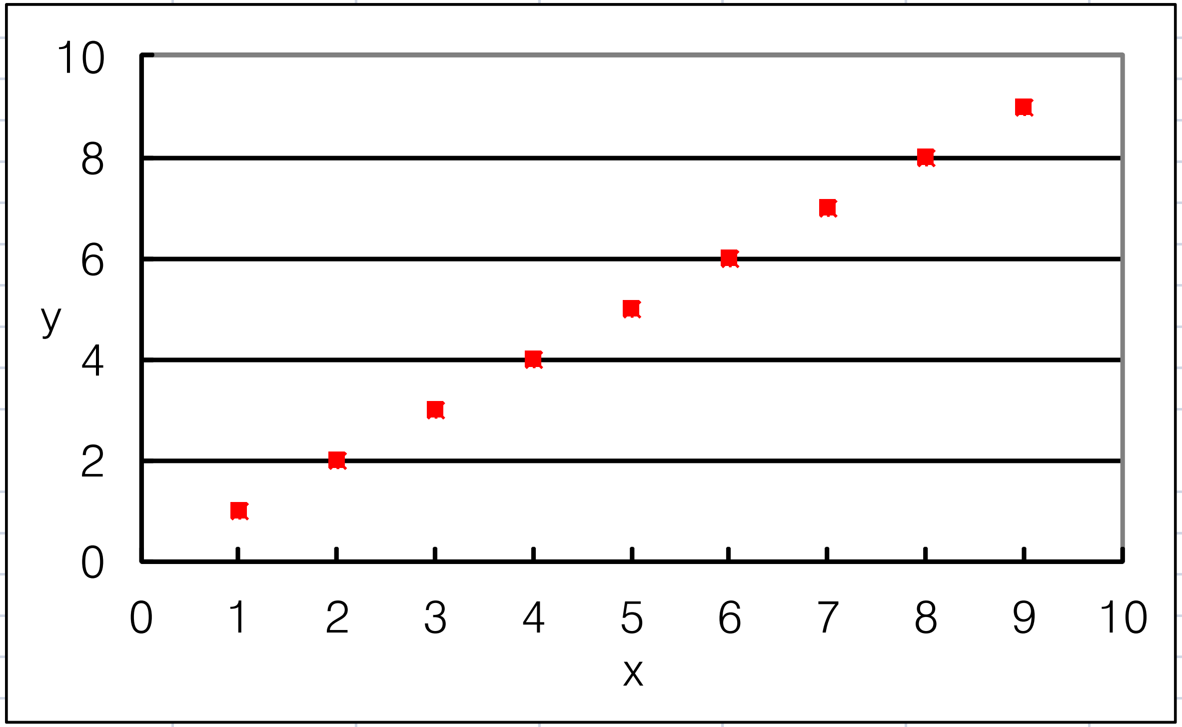
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

공분산을 정규화(normalize) 하기 위해 두 변수의 표준편차로 나누어 줌 → 상관계수는 -1에서 1까지의 범위를 가지게 됨

표본상관계수 r의 성질

< 예제1 >

x	y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
r= 1	

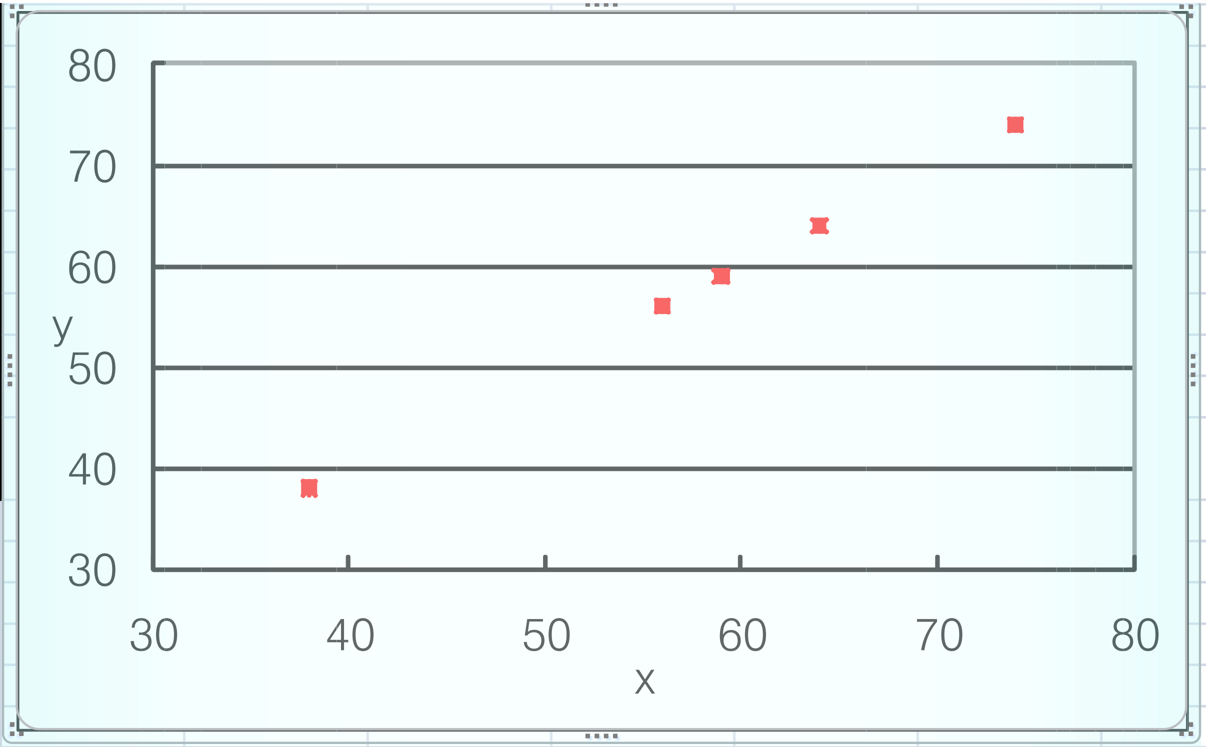


표본상관계수 r의 성질

< 예제2 >

대퇴부	상박부
38	41
56	63
59	70
64	72
74	84

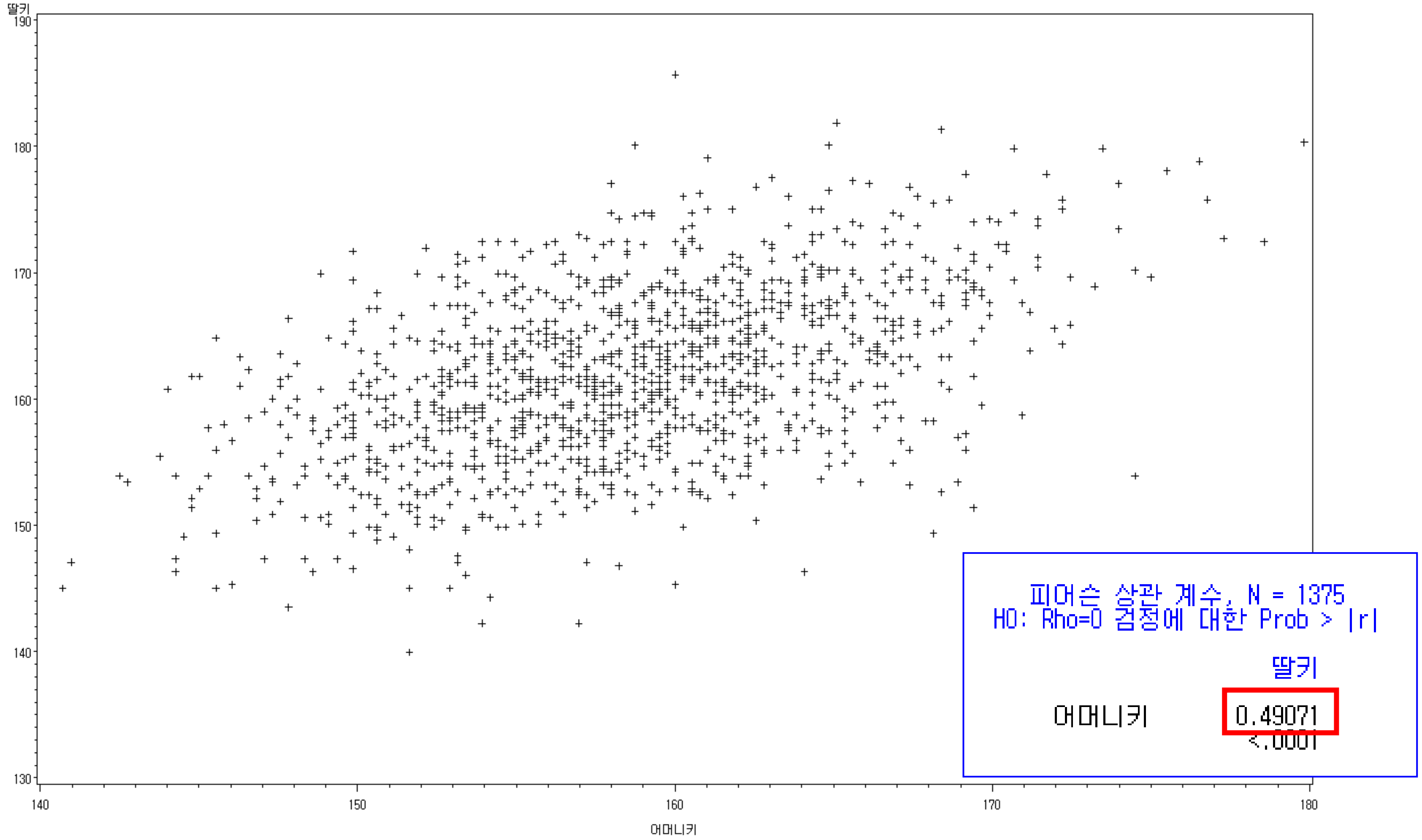
r= 0.99



표본상관계수 r의 성질

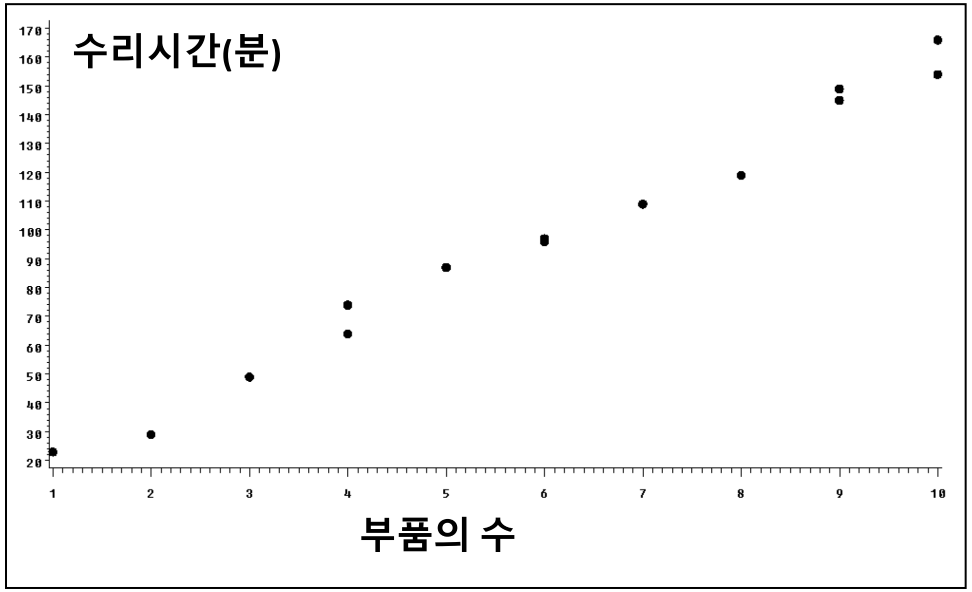
어머니(65미만)와 딸(18세 이상)의 신장(cm)

표본상관계수(sample coefficient of correlation)



표본상관계수 r의 성질

부품 수(x)	1	2	3	4	4	5	6	6	7	8	9	9	10	10
수리시간(y) (단위: 분)	23	29	49	64	74	87	96	97	109	119	149	145	154	166



$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \ \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= 0.9937$$

R-Square : 상관계수의 제곱값으로 독립 변수에 의해 설명되는 종속 변수의 분산 비율

R-Squared

- R-Squared : 회귀 모델에서 독립변수가 종속변수를 얼마만큼 잘 설명해주는지를 판단하는 지표 → 설명력

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

$$\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}\right)^2$$

공변량 측정

변동성 측정

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ → 총 변동(Total Sum of Squared)

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ → 잔차 제곱합(Sum of Squared Residuals/Error)

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ → 회귀에 의해 설명된 변동(Explained Sum of Squares)

R-Squared

회귀분석의 구조

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

- SST(Sum of Squares Total) : 종속변수의 전체 변동량 → SST는 모델이 데이터의 변동성을 얼마나 잘 설명하는지 평가하는 데 필요한 컨텍스트를 제공함
- SSE는 가장 적합한 선을 찾기 위해 오차를 최소화하는 과정에서 사용됨
- $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

R-Squared

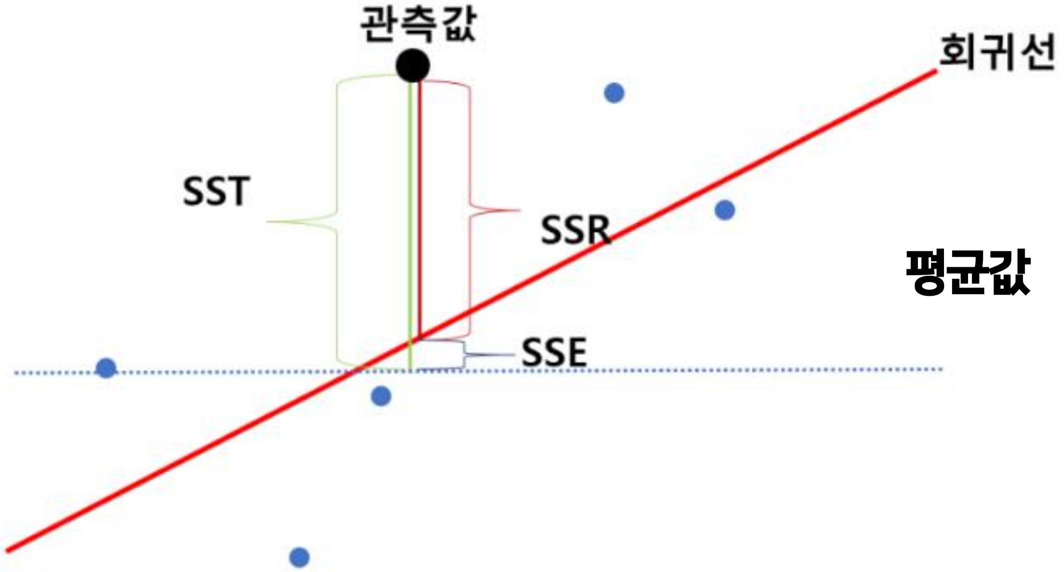
- R-Squared : 회귀 모델에서 독립변수가 종속변수를 얼마만큼 잘 설명해주는지를 판단하는 지표 → 설명력

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

공변량 측정

$$\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \right)^2$$

변동성 측정



R-Squared

- **R-Squared : 회귀 모델에서 독립변수가 종속변수를 얼마만큼 잘 설명해주는지를 판단하는 지표 → 설명력**

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

$y = 11.429 + 1.5357x$

x	y_i	$\hat{y}_i$
10	30	26.8
20	40	42.1
30	50	57.5
40	80	72.9
50	90	88.2
60	100	103.6
70	120	118.9

$\bar{y} = 72.86$

R-Squared

- R-Squared : 회귀 모델에서 독립변수가 종속변수를 얼마만큼 잘 설명해주는지를 판단하는 지표 → 설명력

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

y= 11.429 + 1.5357x

x	y_i	$\hat{y}_i$	$(y_i - \bar{y})$ = Data =	$(\hat{y}_i - \bar{y})$ + Fit +	$(y_i - \hat{y}_i)$ Error
10	30	26.8	-42.86	-46.07	3.21
20	40	42.1	-32.86	-30.72	-2.14
30	50	57.5	-22.86	-15.36	-7.50
40	80	72.9	7.14	0.00	7.14
50	90	88.2	17.14	15.35	1.79
60	100	103.6	27.14	30.71	-3.57
70	120	118.9	47.14	46.07	1.07

$\bar{y}$= 72.86

R-Squared

- **R-Squared → 독립변수의 수가 증가하면 실제로 값이 상승함 즉 결정계수만 가지고 회귀 모델의 유용성을 판단하지 못함**
- **조정된 결정계수(Adjusted R-Squared)**

$$Adjusted\ R^2 = 1 - \frac{SSE \div (n-k-1)}{SST \div (n-1)}$$

n은 표본수 k는 독립변수의 개수 → 자유도를 감안한 방법

R-Squared

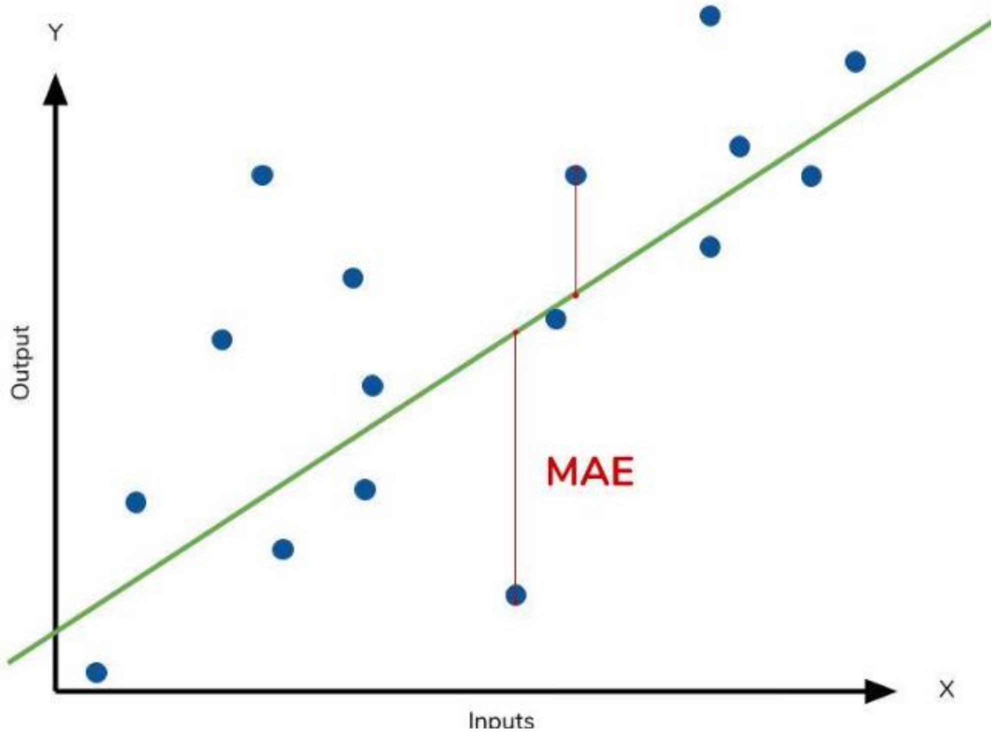
- R-Squared → 보통 0부터 1까지의 값으로 설명되지만 음수가 될 수 있음
- 음수일 경우에는 절편을 포함하지 않는 회귀 모델을 사용할 경우, 잘못된 예측을 할 경우
- 하지만 비율로써 상대적인 설명력을 가지고 있기 때문에 절대적 오차의 크기를 알 수 없음
- 종속변수의 절대적인 오차의 크기를 알기 위해서는 다른 방법들이 필요함

회귀분석으로 알수 있는 것들

- 모형의 적합도 : 모형이 데이터에 얼마나 잘 맞는가?/모형이 얼마나 데이터를 잘 설명하는가?
- 회귀계수 : 독립변수의 변화가 종속변수를 얼마나 변화시키는가?

MAE(Mean Absolute Error) : 평균절대오차

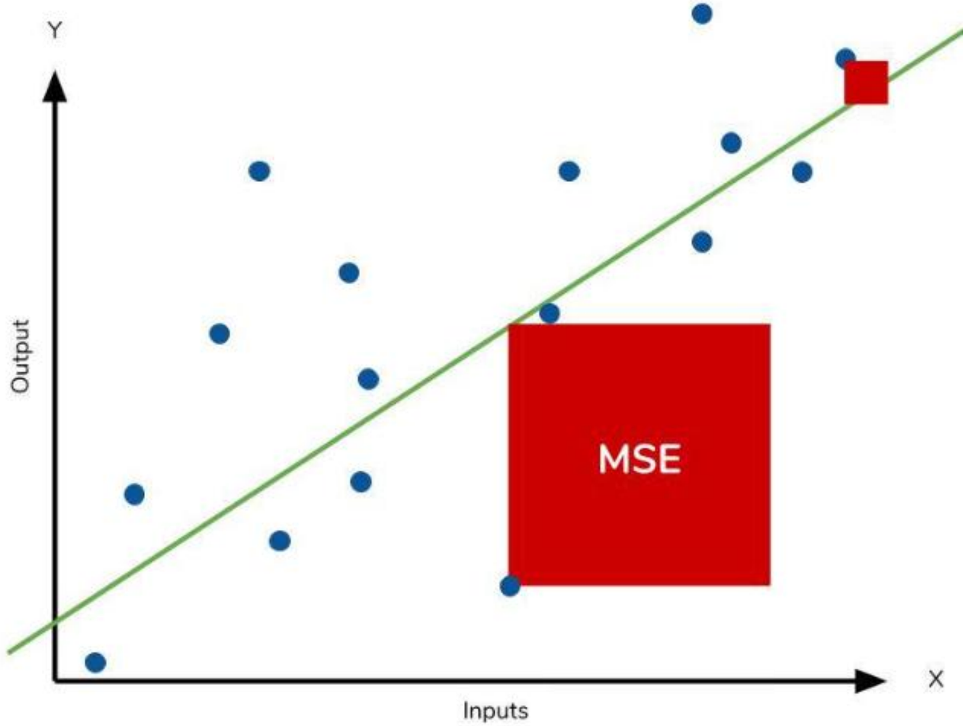
- 모델의 예측값과 실제값의 차이를 더해 절대값을 취하는 지표
- 절대값을 사용하기 때문에 실제보다 낮은지 큰값인지 알 수 없음



$$MAE = \frac{\sum |y - \hat{y}|}{n}$$

MSE(Mean Squared Error) : 평균제곱오차/RMSE(Root MSE(Mean Squared Error) : 평균 오차

- 실제값과 예측값의 제곱을 통해 차이를 판단함
- 특이값이 존재하면 수치가 많이 늘어난다 ➔특이값에 민감함
- RMSE→오류 지표를 실제 값과 유사한 단위로 변환하여 해석을 쉽게 변환함



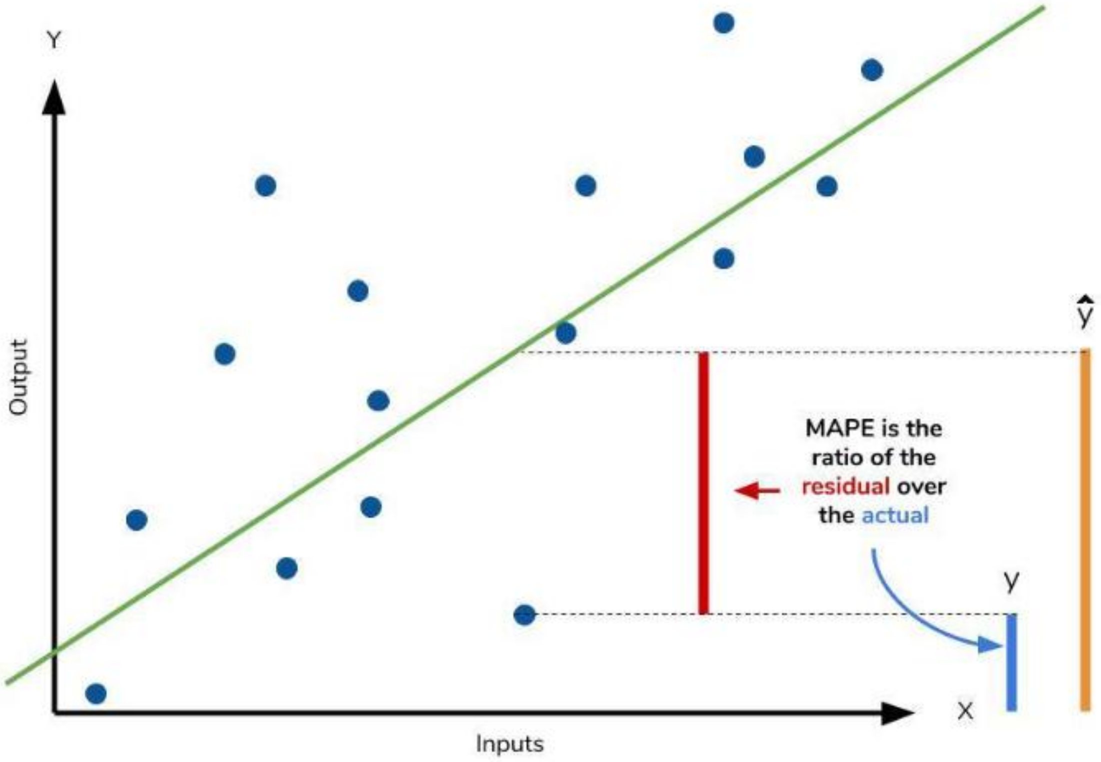
$$MSE = \frac{\sum(y-\hat{y})^2}{n}$$

$$RMSE = \sqrt{\frac{\sum(y-\hat{y})^2}{n}}$$

02 회귀(Regression)

MAPE(Mean Absolute Percentage Error) : 평균 절대 백분오차 비율

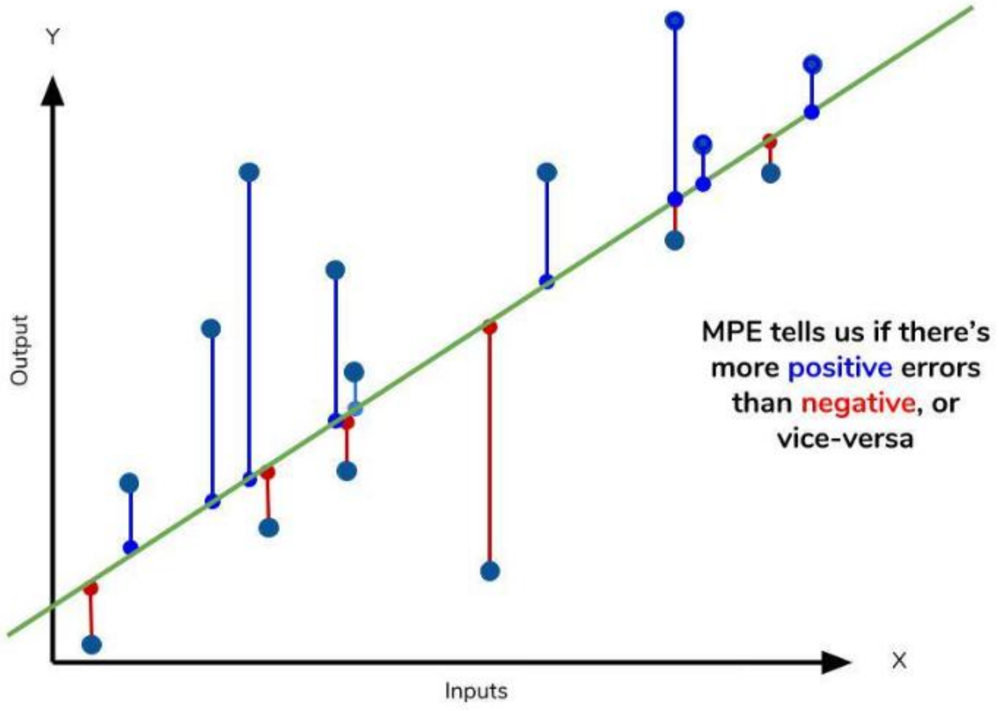
- **MAE를 퍼센트로 변환한 것**
- **MAE와 동일하게 MSE보다 특이값에 영향을 크게 받지 않음**



$$MAPE = \sqrt{\frac{\sum\left|\frac{y-\hat{y}}{y}\right|}{n} * 100\%}$$

MPE(Mean Percentage Error) : 평균 비율 오차

- MAE를 퍼센트로 변환한 것
- MAE와 동일하게 MSE보다 특이값에 영향을 크게 받지 않음



$$MPE = \frac{\sum(y - \hat{y})}{n} * 100\%$$

MPE(Mean Percentage Error) : 평균 비율 오차

- R-Squred vs RMSE …?
- 예측 값과 실제 값 간의 차이를 측정하여 모델이 데이터에 얼마나 잘 맞는지를 판단하는 것 → 예측력(모형의 안정성)
- R-Squred는 모델의 독립 변수에 의해 설명되는 종속 변수의 분산 비율을 측정함 → 설명력
- 종속변수의 변화를 잘 설명 했는지를 판단 → 설명력(종속변수의 변동성을 독립변수들이 얼마나 잘 설명했는지)

종류	Full Name	Residuals Operation? (잔차 계산)	Robust To Outliers? (이상치 영향)
MAE	Mean Absolute Error (평균절대오차)	Absolute Value (절대값)	Yes
MSE	Mean Squared Error (평균제곱오차)	Square (제곱값)	No
RMSE	Root Mean Squared Error (평균오차)	Square (제곱값)	No
MAPE	Mean Absolute Percentage Error (평균 절대 백분 오차 비율)	Absolute Value (절대값)	Yes
MPE	Mean Percentage Error (평균 비율 오차)	N/A	Yes

실습

```
#데이터 불러오기
train_data <- read.csv("simple_train_data.csv")
test_data <- read.csv("simple_test_data.csv")

#회귀분석 모델
model <- lm(upper_arm ~ thigh, data = train_data)
summary(model)

#모델 저장
saveRDS(model, "regression_model.rds")

#모델 불러오기
loaded_model <- readRDS("regression_model.rds")

#불러온 모델을 활용한 예측
predicted <- predict(loaded_model, newdata = test_data)
head(predicted)
```

실습

```r
ggplot(train_data, aes(x = thigh, y = upper_arm)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # 회귀선 (표준오차 리본 제거)
  labs(title = "대퇴부 vs 상박부 회귀분석",
       x = "대퇴부 둘레 (cm)",
       y = "상박부 둘레 (cm)")


# 예측 결과와 실제값 비교
results <- data.frame(Actual = test_data$upper_arm, Predicted = predicted)

#RMSE 계산
rmse <- sqrt(mean((results$Actual - results$Predicted)^2))
print(paste("RMSE:", round(rmse, 3)))
```

실습

```
#회귀분석 모델(절편의 고정)
model <- lm(upper_arm ~ 0 + thigh, data = train_data)
summary(model)
```