통계기초 3

미래를 예측하는 선형 회귀

차이를 설명하고 예측하는 회귀모형

1. 예측의 개념 이해

2. 산점도와 추세선

3. 선형 회귀 모형의 개념

4. 회귀 계수의 계산과 예측

변수의 관계를 활용한 예측

예측(prediction)

- 주어진 정보를 활용해서 불확실한 미래를 대비
- 미래의 가능성이나 불확실성을 숫자로 계산해서 활용
- 결국 변수의 관계를 활용

관심변수와 설명변수

관심변수(=반응변수, 종속변수)

-예측의 대상이 되는 변수

-관심변수에 관측치간 차이가 존재

-다양한 방법으로 관심변수 속 차이를 확인 가능

설명변수(=독립변수)

-관심변수 속 차이를 설명할 수 있는 변수

-관심변수와 설명변수의 관계를 확인하여 예측에 활용

수치형 관심변수와 조건부 평균

한 수치형 변수를 관심변수로 지정

- 한 수치형 변수 요약 : 평균 계산
- 평균을 중심으로 관측치들이 흩어진 차이가 존재

조건부 평균(conditional mean)

- 특정 설명변수 조건과 일치하는 부분 관측치로 계산된 평균
- 범주형 설명 변수를 활용 : 그룹별 평균
- 수치형 설명 변수를 활용 : 선형 회귀

범주형 설명변수를 활용한 예측

한 범주형 변수를 설명변수로 지정

–수준에 따라 그룹별 평균을 계산

–각 관측치의 수준을 파악해 예측에 활용 가능

- 예제) 고객 1,000명의 거주지역과 방문횟수의 요약

거주지역	가	나	다	라	전체
평균 방문횟수	13.9	12.9	15.2	12.8	13.8

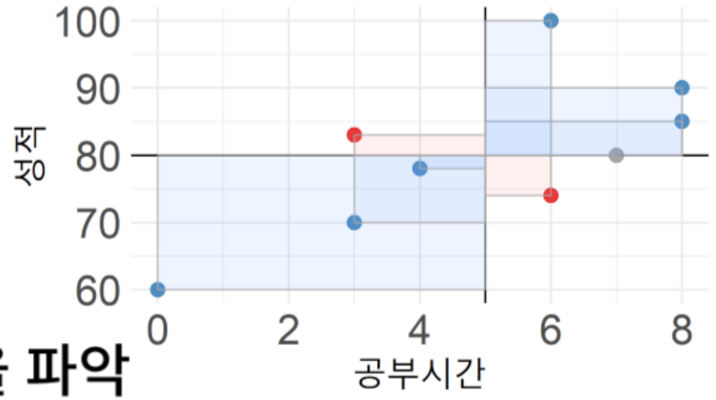
- 신규고객의 방문횟수 예측 : 전체 평균 13.8 혹은 거주지역에 따른 예측 가능

수치형 설명변수를 활용한 예측 전략

①수치형 설명변수의 구간화

–수치형 설명변수의 구간화를 통한 그룹별 평균 계산 가능

- 예제) 공부시간대별 성적차이 계산



②산점도와 상관계수의 활용

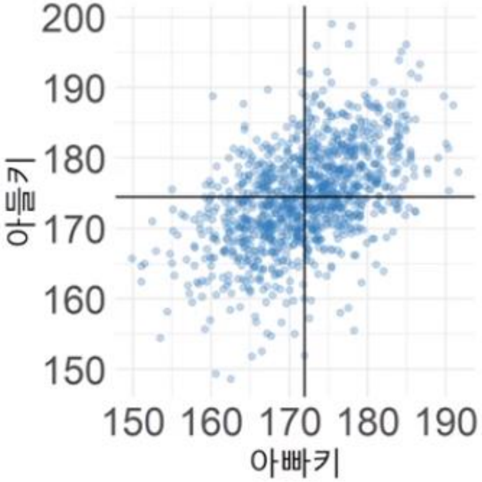
–산점도 : 사분면에서 관측치의 흩어진 패턴을 파악

–상관계수 : 두 수치형 변수의 관계를 -1부터 1사이의 숫자로 표현

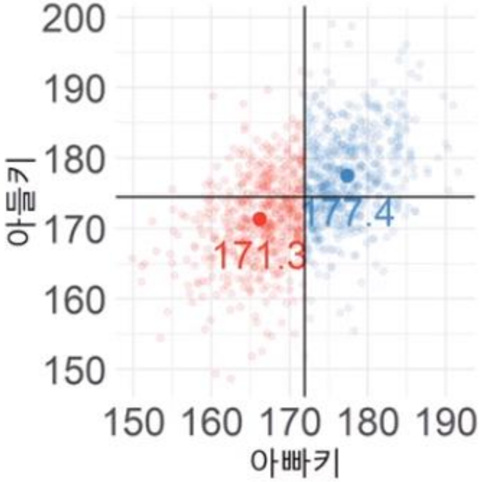
수치형 변수를 활용한 예측 예제

예제) 아빠키-아들키 데이터의 산점도와 구간화를 활용한 예측

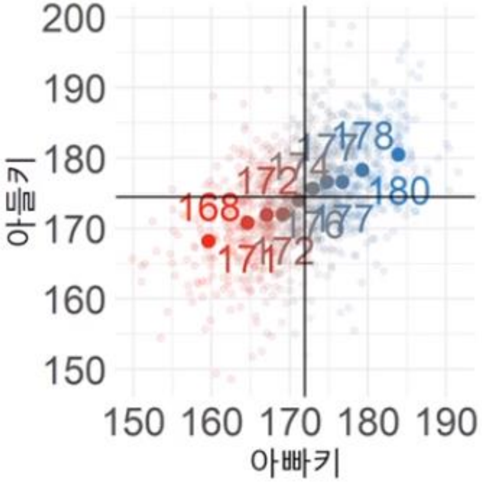
· 두 변수의 산점도



· 두 그룹으로 나눈 예측



· 10개 그룹으로 나눈 예측



· 그룹을 잘게 쪼개는 개념을 확장하면 직선으로 예측 가능

일차함수와 추세선

①일차함수

- 두 변수 X와 Y의 정비례 관계를 가정

$$Y = a + bX$$

- “X가 1씩 커질 때 마다 Y는 b만큼 비례해서 변화”
- 일차함수는 공간에서 직선으로 표현

②추세선

- 일차함수를 활용하여 두 수치형 변수의 관계를 설명가능

선형 회귀 모형 이해를 위한 표기법 정의

표기법 정의

- Y : 수치형 관심변수
- X : 수치형 설명변수
- β_0, β_1 : 회귀 계수(regression coefficient)
- ε : 오차(error), 랜덤으로 정해지는 설명할 수 없는 부분

선형 회귀 모형의 개념

단순 선형 회귀(simple linear regression)

- 수치형 관심변수를 수치형 설명변수의 정비례로 설명하는 모형

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

- “X가 1씩 커질 때 마다 Y는 b만큼 비례해서 변화”
- “단, Y에는 X로는 설명할 수 없는 오차 ε가 존재”

- 설명변수 X와 회귀 계수를 활용해서 관심변수 Y를 예측 가능

선형회귀모형의 적합

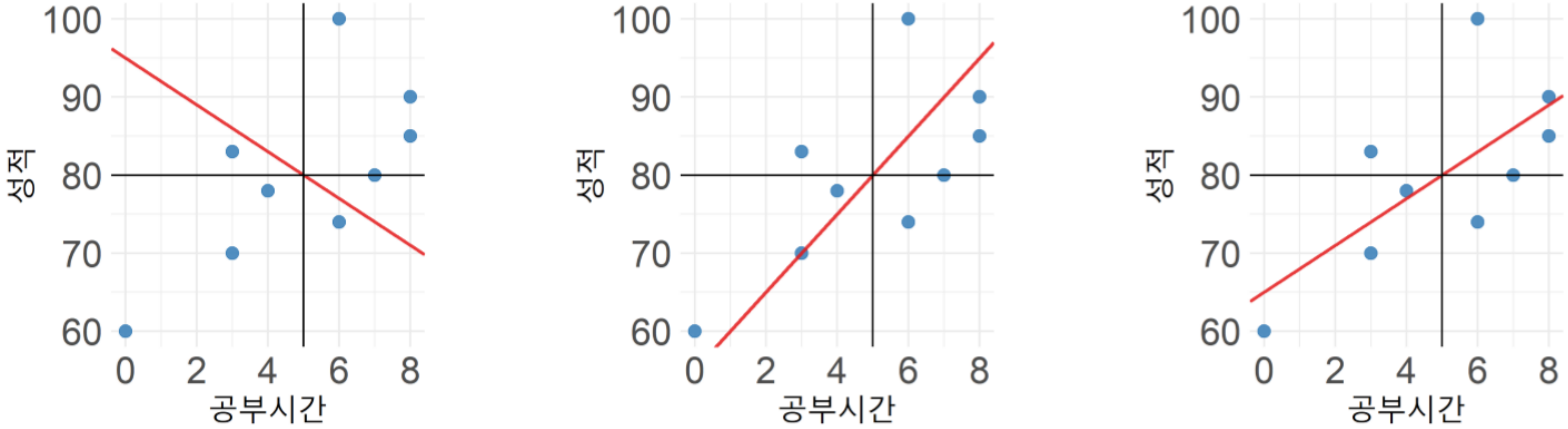
회귀 모형의 적합

- 두 변수 X와 Y의 관계식을 확인하는 과정
- 데이터를 활용하여 회귀 직선 $Y = \beta_0 + \beta_1 X$에서 가장 적절한 β_0과 β_1을 계산하는 과정
- 결국 산점도에 가장 적절한 추세선을 긋는 문제

산점도와 추세선

예제) 공부시간과 성적의 산점도와 세가지 추세선

- 가장 합리적인 추세선은?



통계적으로 합리적인 추세선의 조건

최소 제곱법(least squares method)을 활용

① X가 평균정도 일 때는 Y도 평균정도로 예측

：따라서 모든 회귀직선은 무게 중심 $(\bar{X}, \bar{Y})$을 지남

② 추정된 회귀계수 β_0, β_1와 X를 활용한
예측 값과 실제 값 Y의 전반적인 차이가 적음

：최적의 직선의 기울기 β_1을 데이터로부터 계산

추정된 회귀 계수와 상관계수의 관계

통계학자가 계산한 최적의 회귀계수 추정값 $(\hat{\beta}_0, \hat{\beta}_1)$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r_{XY}\frac{s_Y}{s_X}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

- 회귀 직선의 기울기 β_1는 두 변수의 상관계수에 비례

회귀 계수를 활용한 예측

$$\hat{\beta}_0 + \hat{\beta}_1 X$$

회귀 계수와 추세선 예제

예제) 공부시간과 성적의 회귀 계수 추정

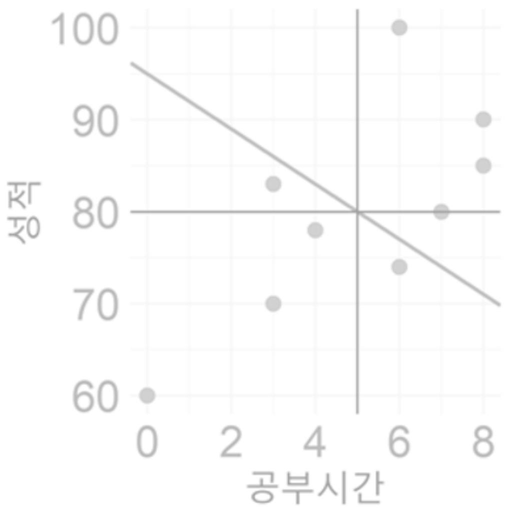
$$\hat{\beta}_1 = 3,\ \hat{\beta}_0 = 80 - 3 \times 5 = 65$$

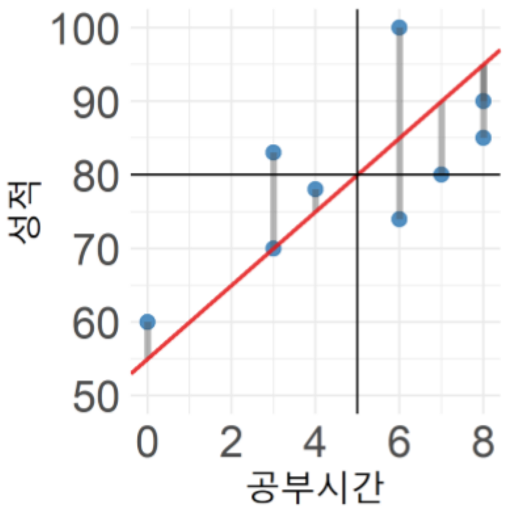
$$\Rightarrow \text{성적} = 65 + 3 \times \text{공부시간}$$

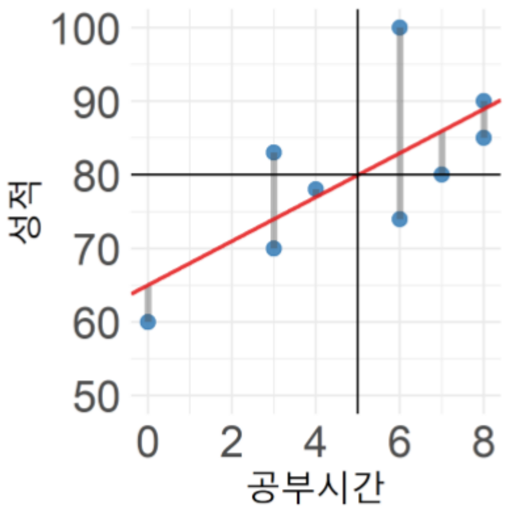
회귀 계수와 추세선 비교

예제) 공부시간과 성적 회귀 직선 : 성적 $= 65 + 3 \times$ 공부시간

- 기울기가 더 큰 것보다 예측 값과 실제 값의 전반적인 차이가 더 적음







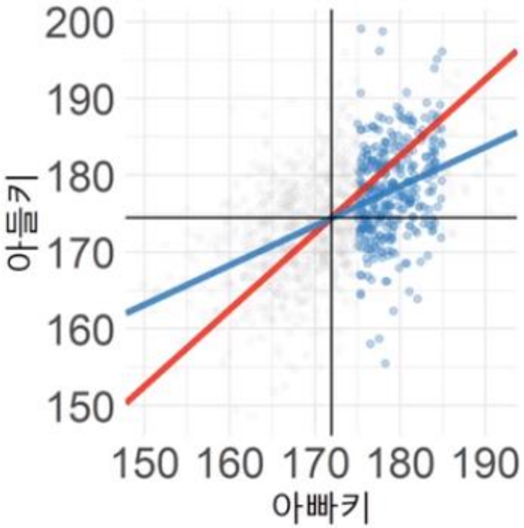
산점도에서 살펴보는 회귀의 의미

예제)아빠키-아들키 데이터의 산점도

- 회귀 모형을 활용한 추세선 추가(파란색)

$$\text{아들키} = 86.07 + 0.514 \times \text{아빠키}$$

  - 아빠키가 180cm 일 때 : 아들키 예측 값 = 178.6cm



회귀(regression)

- X가 꽤 커도 Y는 생각보다 작게 예측
- X가 꽤 작아도 Y는 생각보다 크게 예측
- 예측된 Y값이 평균(중심)으로 당겨지는 효과 → “Regression”

더 나아간 모형

더 많은 설명 변수를 활용한 회귀모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p + \varepsilon$$

- p개의 설명변수를 활용가능
- 데이터로 적절한 회귀계수를 추정하고 예측에 활용

$$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

다양한 모형의 활용

- 의사결정나무 모형 등 활용 가능