# Indicator Variable

# 변수의 종류

- 양적변수 (Quantitative Variable)
  - 어떤 구간에 속하는 모든 값을 관측값으로 취할 수 있는 변수
  - 온도, 압력, 습도, 무게, 거리 등
  - 양적으로 비교 가능
- 질적 (Qualitative) 변수 또는 범주형 (Categorical) 변수
  - 양적으로 비교 불가능
  - 성별(남여), 신용상태(좋음, 나쁨), 보험의 종류(생명,손해 보험)

# Indicator Variable

- 가변수(Dummy variable) 또는 이진형 변수 (Binary Variable)
  - 질적변수가 두 개의 범주로 이루어진 경우 “0” 또는 “1” 로 설정
  - 설명변수 또는 반응변수로 사용 가능

# Example

- 데이터

| 번호 | $y$ | $x_1$ | $x_2$ | 번호 | $y$ | $x_1$ | $x_2$ |
|---|---|---|---|---|---|---|---|
| 1 | 17 | 151 | 남자 | 11 | 28 | 164 | 여자 |
| 2 | 26 | 92 | 남자 | 12 | 15 | 272 | 여자 |
| 3 | 21 | 175 | 남자 | 13 | 11 | 295 | 여자 |
| 4 | 30 | 31 | 남자 | 14 | 38 | 68 | 여자 |
| 5 | 22 | 104 | 남자 | 15 | 31 | 85 | 여자 |
| 6 | 1 | 277 | 남자 | 16 | 21 | 224 | 여자 |
| 7 | 12 | 210 | 남자 | 17 | 20 | 166 | 여자 |
| 8 | 19 | 120 | 남자 | 18 | 13 | 305 | 여자 |
| 9 | 4 | 290 | 남자 | 19 | 30 | 124 | 여자 |
| 10 | 16 | 238 | 남자 | 20 | 14 | 246 | 여자 |

# Example

- Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \;\; i = 1, 2, \ldots, n$$

- $y$ : 새로운 프로그램을 익히는 데 소요된 시간
- $x_1$ : 교육 시작 전 실시한 적성검사 시험의 점수 (500점 만점)
- $x_2$ : 성별

$$x_{i2} = \begin{cases} 0, & \text{Male} \\ 1, & \text{Female} \end{cases}$$

# Example

- 남자인 경우 : $y = \beta_0 + \beta_1 x_1 + \epsilon$

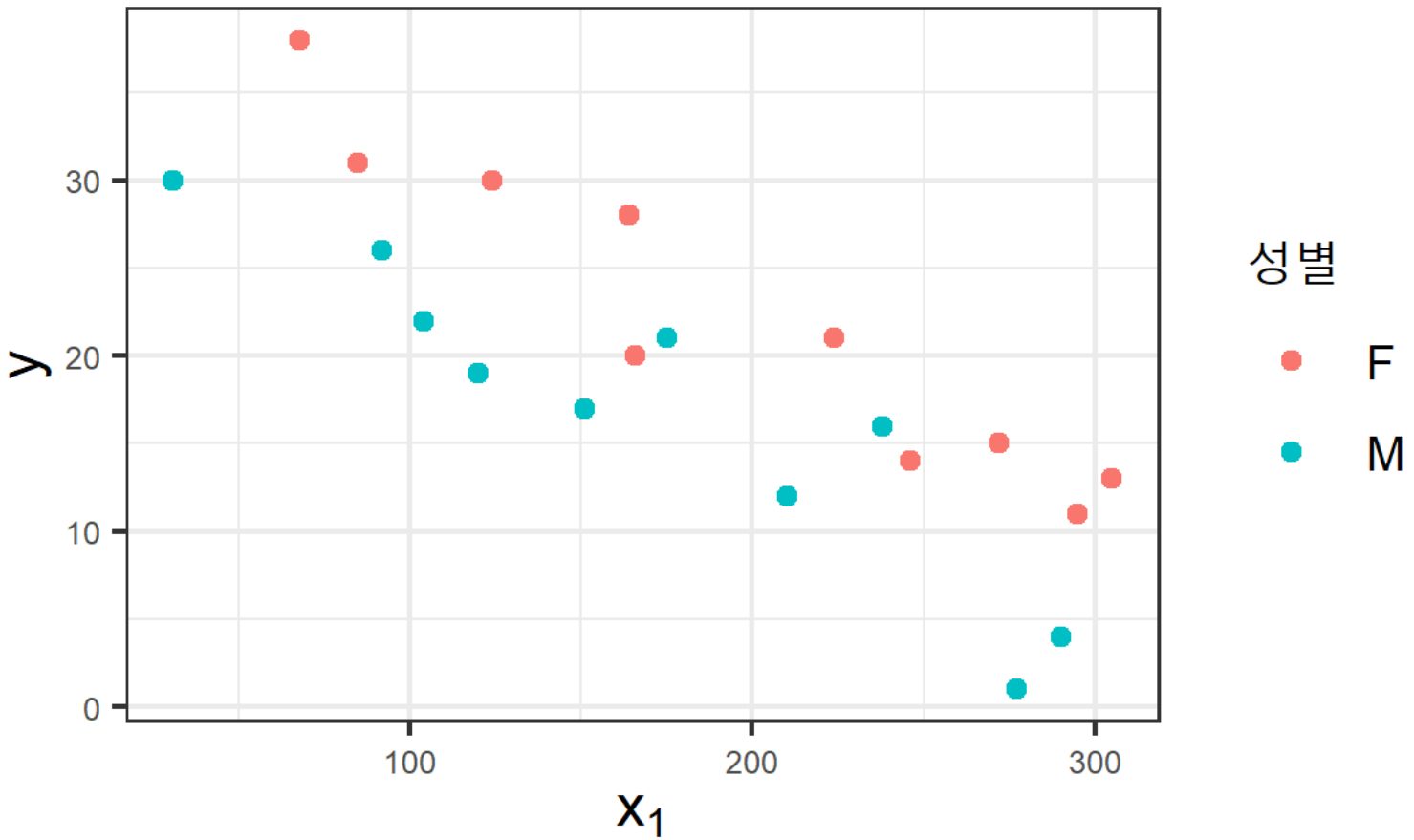
$$E(y|M) = \beta_0 + \beta_1 x_1$$

- 여자인 경우 : $y = \beta_0 + \beta_1 x_1 + \beta_2 + \epsilon$

$$E(y|F) = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1$$

- 기울기는 성별로 차이가 없지만, 평균시간은 $\beta_2$ 만큼 차이가 남

# Example



# EXample

- 회귀계수의 추정 (최소제곱추정량)

$$\boldsymbol{y} = \begin{pmatrix} 17 \\ 26 \\ \vdots \\ 14 \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & 151 & 0 \\ 1 & 92 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 246 & 1 \end{pmatrix} \Rightarrow \ \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 33.8349 \\ -0.1009 \\ 7.9340 \end{pmatrix}$$

  - 회귀직선

$$\hat{y} = 33.8349 - 0.1009 x_1 + 7.9340 x_2$$

# EXample

- 회귀계수의 추정 (최소제곱추정량)
  - 남자인 경우

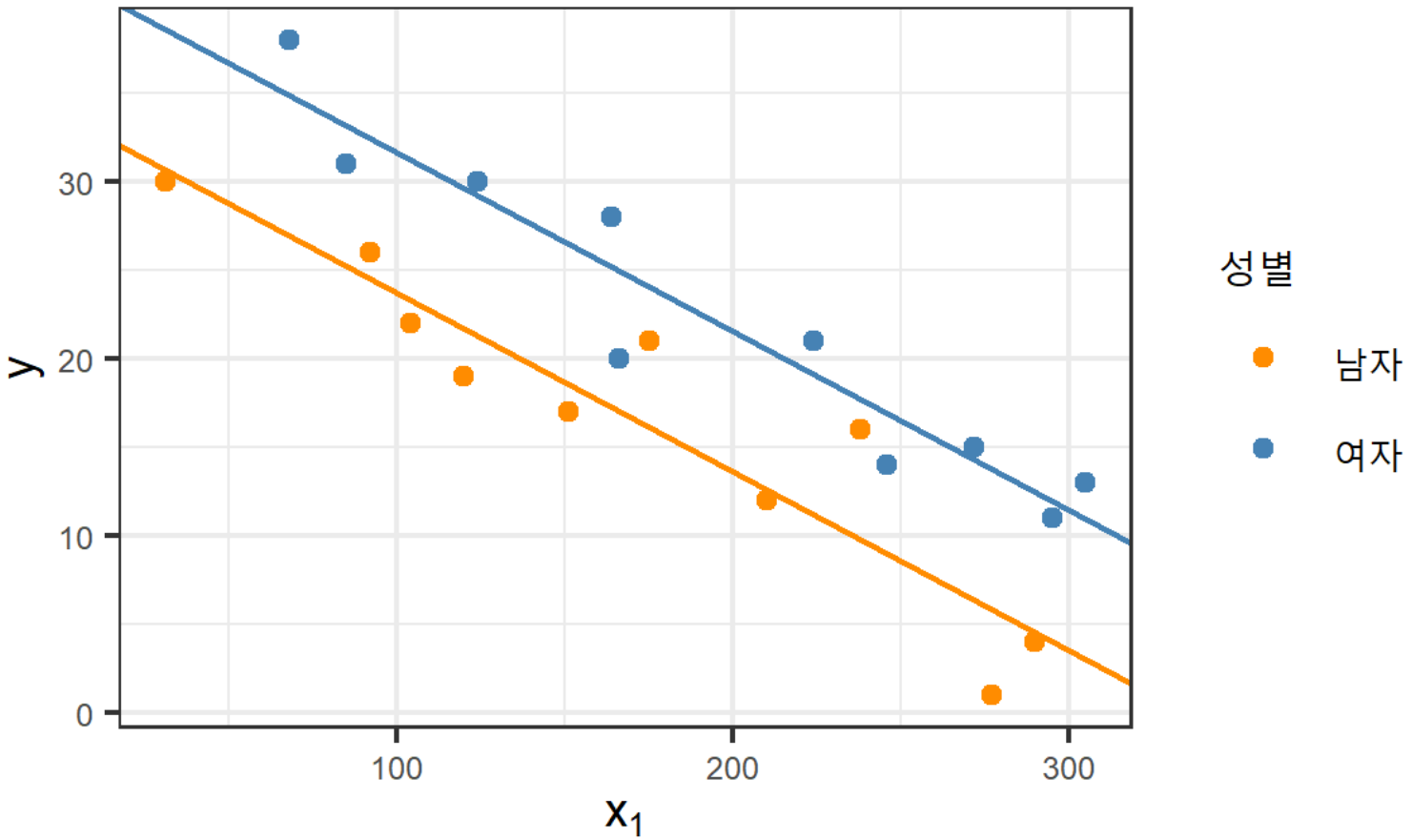
$$\widehat{E(y|M)} = \hat{y} = 33.8349 - 0.1009x_1$$

  - 여자인 경우

$$\widehat{E(y|F)} = \hat{y} = 33.8349 - 0.1009x_1 + 7.9340$$

$$= 41.7689 - 0.1009x_1$$

# Example



# Example

- 분산분석표

| 요인 | 제곱합 | 자유도(df) | 평균제곱(MS) | $F_0$ | $F_{0.05}(2, 17)$ |
|---|---|---|---|---|---|
| 회귀 | $SSR = 1477.14$ | 2 | $MSR = 738.57$ | 75.75 | 3.59 |
| 잔차 | $SSE = 165.81$ | 17 | $MSE = 9.75$ | | |
| 계 | $SST$ | 19 | | | |

# Example

- 유의성검정
  - $H_0 : \beta_2 = 0$
    - t-test
    - Extra sum of squared method (partial F-test

# 교호작용

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

- 평균반응 : $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
- 남자인 경우 : $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

$$E(y|M) = \beta_0 + \beta_1 x_1$$

- 여자인 경우 $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1} + \epsilon_i$

$$E(y|F) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$$

# 교호작용 - EXample

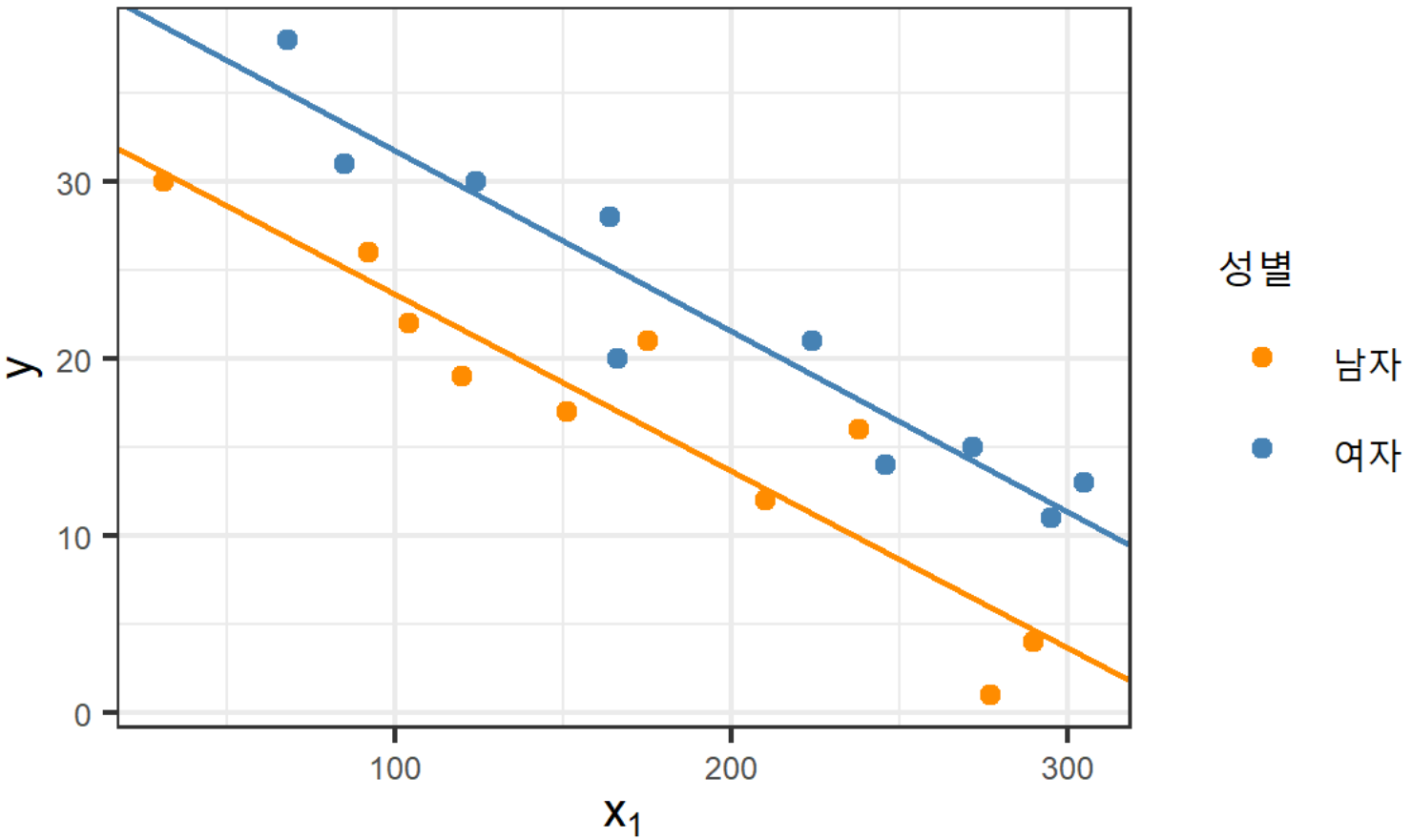
- 회귀계수의 추정 (최소제곱추정량)

$$\boldsymbol{X} = \begin{pmatrix} 1 & 151 & 0 & 0 \\ 1 & 92 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 246 & 1 & 246 \end{pmatrix} \Rightarrow \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 33.6561 \\ -0.0999 \\ 8.3135 \\ -0.0021 \end{pmatrix}$$

  - 회귀직선

$$\hat{y} = 33.6561 - 0.0999 x_1 + 8.3135 x_2 - 0.0021 x_1 x_2$$

# 교호작용 - Example



# 교호작용 - EXample

- 가설검정 : $H_0 : \beta_3 = 0$
  - 검정통계량 : $t = \dfrac{\hat{\beta}_3}{\widehat{s.e.}(\hat{\beta}_3)} \sim_{H_0} t(n-p-1) = t(16)$
  - 검정통계량의 관측값 : $t_0 = \dfrac{-0.0021}{0.017766} = -0.118$
  - 기각역 : $|t_0| \geq t_{\alpha/2}(n-p-1) = t_{0.025}(16) = 2.120$
  - 결론 ; 기각역에 속하지 않으므로 귀무가설을 기각할 수 없다. 즉 교호작용의 효과가 거의 없다고 할 수 있다.

# 3개 이상의 범주를 갖는 질적변수

- 예) 시험성적($x_1$), 학력 : 고졸($E_1$), 대졸($E_2$), 대학원이상($E_3$)

$$x_2 = \begin{cases} 1, & E_1 \\ 0, & o.w. \end{cases}, \; x_3 = \begin{cases} 1, & E_2 \\ 0, & o.w. \end{cases}, \; x_4 = \begin{cases} 1, & E_3 \\ 0, & o.w. \end{cases}$$

- model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

- $rank(X) < 5 \Rightarrow$ $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ 가 존재하지 않는다. $\Rightarrow$ $x_4$ 제외

# 3개 이상의 범주를 갖는 질적변수

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

- 평균반응

$$\begin{cases} E_1 : E(y|E_1) = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1 \\ E_2 : E(y|E_2) = \beta_0 + \beta_1 x_1 + \beta_3 = (\beta_0 + \beta_3) + \beta_1 x_1 \\ E_3 : E(y|E_3) = \beta_0 + \beta_1 x_1 \end{cases}$$

# 3개 이상의 범주를 갖는 질적변수

- 교호작용포함

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \epsilon_i$$

  - 평균반응

$$\begin{cases} E_1 : E(y|E_1) & = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_4 x_1 \\ & = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 \\ E_2 : E(y|E_2) & = \beta_0 + \beta_1 x_1 + \beta_3 + \beta_5 x_1 \\ & = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 \\ E_3 : E(y|E_3) & = \beta_0 + \beta_1 x_1 \end{cases}$$

# 구간별 회귀분석

- Example
  - 예) 단가($y$), 주문량 ($x$) : 주문량이 기준값($x_w$)보다 커지면 단가가 갑자기 작아짐
  - model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - x_w) x_{i2} + \epsilon_i$$

    - $x_{i1}$ : 주문량
    - $x_{i2} = \begin{cases} 0, & x_{i1} < x_w \\ 1, & x_{i1} \geq x_w \end{cases}$

# 구간별 회귀분석

- Example
  - 평균반응

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - x_w) x_2$$

  - 주문량별 평균반응

$$E(y|x_1 < x_w) = \beta_0 + \beta_1 x_1$$

$$\begin{aligned} E(y|x_1 \geq x_w) &= \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - x_w) \\ &= (\beta_0 - x_w \beta_2) + (\beta_1 + \beta_2) x_1 \end{aligned}$$

    - $\beta_2$ : 기울기의 차이

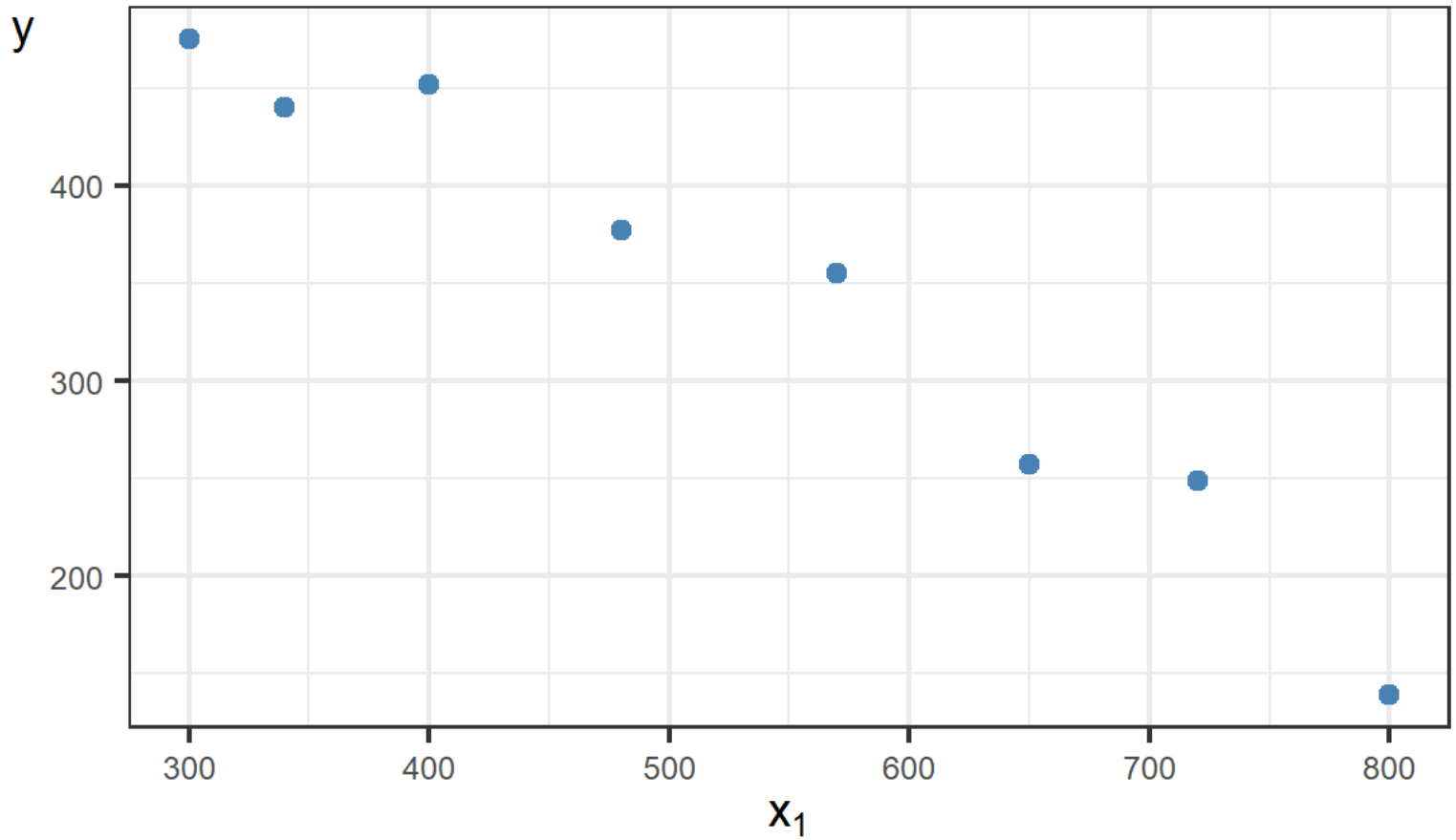
# 구간별 회귀분석

Table: 태블릿 주문량 데이터

| 기업체번호 | 주문량($x_1$) | 단가(y) |
|---|---|---|
| 1 | 480 | 377 |
| 2 | 720 | 249 |
| 3 | 570 | 355 |
| 4 | 300 | 475 |
| 5 | 800 | 139 |
| 6 | 400 | 452 |
| 7 | 340 | 440 |
| 8 | 650 | 257 |

# 구간별 회귀분석

- Example



# 구간별 회귀분석

- Example

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 500) x_{i2} + \epsilon_i, \ i = 1, \ldots, 8$$

$$x_{i2} = \begin{cases} 0, & x_{i1} < 500 \\ 1, & x_{i1} \geq 500 \end{cases}$$

$$\boldsymbol{y} = \begin{pmatrix} 377 \\ 249 \\ \vdots \\ 257 \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & 480 & 0 \\ 1 & 720 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 650 & 150 \end{pmatrix} \Rightarrow \ \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 589.5447 \\ -0.3954 \\ -0.3893 \end{pmatrix}$$

# 구간별 회귀분석

- Example
  - 추정된 회귀직선

$$\hat{y} = 589.5447 - 0.3954x_1 - 0.3893(x_1 - 500)x_2$$

    - $x_1 < 500$ : 주문량 1증가 $\Rightarrow$ 단가 0.3954 감소
    - $x_1 \geq 500$ : 주문량 1증가 $\Rightarrow$ 단가 0.3954+0.3894=0.7848감소

# 구간별 회귀분석

- 구간이 3개인 경우
  - 예) $x_w = 500, 700$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 500) x_{i2} + \beta_3 (x_{i1} - 700) x_{i3} + \epsilon_i$$

    - $x_1$ : 주문량

$$x_2 = \begin{cases} 0, & x_1 < 500 \\ 1, & x_1 \geq 500 \end{cases}, \quad x_3 = \begin{cases} 0, & x_1 < 700 \\ 1, & x_1 \geq 700 \end{cases}$$