9/5(월) 고급회귀분석

- 중간고사 이후 관련 내용 가지고 토론/실습
- 이론 7-80%, 실습 2-30% (R-studio)
- 데이터를 정하고 수업시간에 배운 내용을 적용해보기 → 기말과제 or 기말시험
- 질문: 메일, 수업시간
- 선수: 선형대수, 일반통계 (추정량, 신뢰구간, 가설검정)
- 교재: 회귀분석 (모 저독 모망)

Simple Linear Regression

두 변수 사이의 관계

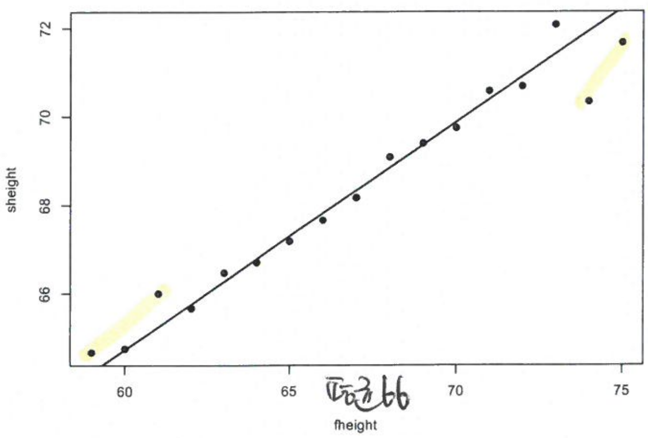
- 대략적 파악 : 산점도(scatter plot)
- 상관분석(correlation analysis)
  - 두 변수 사이의 상관관계 분석
  - 확률변수 $X, Y \to \rho = \mathrm{Corr}(X, Y)$ - 직선적인 관련성 파악
    상관계수 (선형관계)
- 회귀분석(regression analysis)
  - 두 변수 사이의 함수관계를 분석
  - x : 독립변수 또는 설명변수, Y : 종속변수 또는 반응변수, 목표변수
    보조변수
    $$Y = f(x) + \varepsilon,\ \varepsilon : \text{오차항} \to f(x)?$$
  - 단순선형회귀분석 - 직선관계를 모형으로 분석
    $$(f(x) = a + bx)$$
  - 중회귀분석 - 두 개 이상의 설명변수 사용
    $$(f(x) = a + b_1x_1 + \cdots + b_kx_k)$$

확률밀도함수 … 질량함수

X: 분포 : pdf (pmf)
E(X). Var(X)

X, Y → pdf 결합확률분포
joint
E(X), E(Y), Var(X), Var(Y)
: cov(X, Y)
⇓
corr(X, Y)

Example



평균 68~69

평균 66

기울기가 45°는 아니고
더완만함

① f↑ ⇒ s↑ 직선관계

② s ⇒ 평균으로 되돌아가려는 성질

Figure: 아버지의 키(fheight)와 아들 키(sheight)간의 회귀직선

회귀분석의 기본개념

n: sample size (표본크기)

- 자료구조

- 자료구조 : $(x_1, Y_1), ..., (x_n, Y_n)$
- $(x_1, \ldots, x_n)$: 설명변수(explanatory variable)(또는 독립변수) ; 상수(숫자)
  두 변수가 있을 때, 다른 한 변수에 영향을 주는 변수
- $(Y_1, \ldots, Y_n)$: 반응변수(response variable)(또는 종속변수) 목표변수 (Target)
  두 변수가 있을 때, 다른 한 변수에 영향을 받는 변수
- 관측값 : $(x_1, y_1), \ldots, (x_n, y_n)$

Example

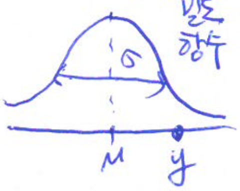
n 상점번호	x 광고료	y 총판매액	상점번호	광고료	총판매액
1	4	9	6	12	30
2	8	20	7	6	18
3	9	22	8	10	25
4	8	15	9	6	10
5	8	17	10	9	20

Table: 표본상점의 광고료(단위:10만원)와 총판매액(단위:100만원)

기본 가정

X가 주어졌을때 Y의 기대값은 일차함수 관계일 것이다.

pdf 확률 밀도 함수



- Linearity (선형성) : $E(Y|X=x) = \mu_{y\cdot x} = \beta_0 + \beta_1 x$ 일차함수관계
- Homoscedastic (등분산성) : $Var(Y|X=x) = \sigma^2$ ← 분산은 x값과 상관이 없다.
- Normality (정규성) : $Y|X=x \sim N(E(Y|X=x), \sigma^2)$
- Independency (독립성) : ϵ's are mutually independent

μ + 오차 = y

$Y_1, \ldots, Y_n$ 서로독립

Example

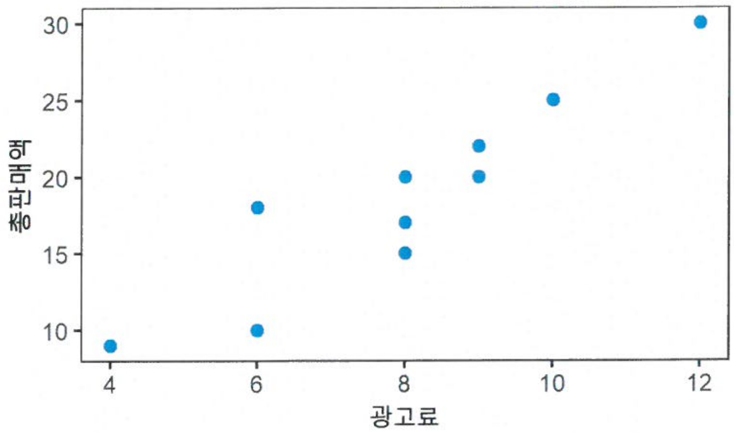


Figure: 광고료와 총판매액의 산점도

산점도를 통해 선형관계 파악
→선형회귀분석 이용

단순선형회귀 모형

IID (independent)

$\epsilon_i \sim N(0,\sigma^2)$

- Model

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

$y|X=x \sim N(\ ,\sigma^2)$

$\epsilon_i = y_i - (\beta_0 + \beta_i x_i)$

- $(\epsilon_1, \ldots, \epsilon_n)$: 오차항(random error)
  서로 독립이면서 평균이 0, 분산이 σ^2 인 확률 변수
- 회귀계수(regression coefficient) (or 모수, parameter)
  - β_0 : 상수항 또는 절편 (constant coefficient or intercept)
  - β_1 : 기울기 (slope)
- 회귀직선, 회귀선 : $\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x}$
  - $\hat{\beta}_0, \hat{\beta}_1, \hat{y}$: estimate of $\beta_0, \beta_1, E(Y|X=x)$

$X \sim N(\mu_x, \sigma^2)$

$\to aX+b \sim N(a\mu_x + b, a^2\sigma^2)$

a, b 상수

$\epsilon_i \sim N(0,\sigma^2)$

y의 추정값
y hat
y의 기대값에 대한 추정값

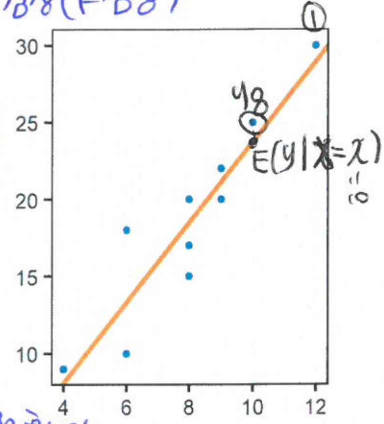
① 산점도
② 모형적합: $\boxed{y=\beta_0+\beta_1 x+\varepsilon_i, \ \varepsilon_i \sim N(0, \sigma^2), \text{iid}}$
1) 추정: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$
추정방법: LSE

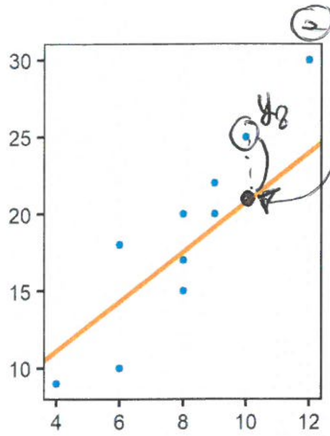
Example - 회귀직선의 비교

2) 추론(가설검정)
ㄱ. 회귀모형의 유의성검정(F검정) $- H_0: \beta_1=0$
ㄴ. 회귀계수의 유의성검정(t검정)

$$\begin{cases} H_0: \beta_0 = \beta_0' (=0) \\ H_1: \beta_1 = \beta_1' (=0) \end{cases}$$

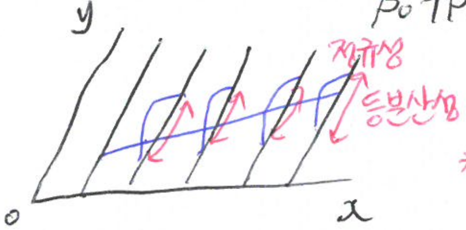
3) 적합도 측정: R^2
4) 회귀진단 { 오차항의 가정확인 등등





$E(y) = \mu_{y.x}$
$= \beta_0 + \beta_1 x$

$y \sim N(\mu_{y.x}, \sigma^2)$
선형성



*독립성

Least Square Estimation (LSE)

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
$\quad\quad\quad\quad \| $
$\quad\quad E(y_i | X = x_i)$

- 오차제곱합

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

- 최소제곱추정량(LSE)

$\hat{\beta}_0, \hat{\beta}_1$ minimize S (β_0, β_1)

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

- Least square fit : $\hat{y} \left(\equiv \widehat{E(Y|X=x)} \right) = \hat{\beta}_0 + \hat{\beta}_1 x$

회귀직선

16page

$\rightarrow \Sigma (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$
$= \Sigma e_i^2$.

절편X

x_i와 y_i는 측정값. 변수는 β_1

$$S = \Sigma(y_i - \beta_1 x_i)^2$$
$$= \Sigma(y_i^2 - 2\beta_1 y_i x_i + \beta_1^2 x_i^2)$$

Least Square Estimation (LSE)

$$= (\Sigma x_i)\beta_1^2 + (\Sigma y_i x_i)\beta_1 + \Sigma y_i^2$$

- 정규방정식 (normal equation)

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$S = \Sigma(y_i - \beta_0 - \beta_1 x_i)^2$$
$$\frac{\partial S}{\partial \beta_0} = \Sigma 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

Least Square Estimation (LSE)

- 최소제곱추정량 = 정규방정식의 해

$$\begin{cases} -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 & (1) \\ -2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 & (2) \end{cases}$$

$$(1)\ \Sigma y_i - \Sigma\hat{\beta}_0 - \Sigma\hat{\beta}_1 x_i = 0$$
$$= \Sigma y_i - n\hat{\beta}_0 - \hat{\beta}_1\Sigma x_i = 0$$
$$(2)\ \Sigma x_i y_i - \Sigma\hat{\beta}_0 x_i - \Sigma\hat{\beta}_1 x_i^2$$
$$= \Sigma x_i y_i - \hat{\beta}_0\Sigma x_i - \hat{\beta}_1\Sigma x_i^2 = 0$$

$$\Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i \\ \hat{\beta}_0\sum_{i=1}^{n}x_i + \hat{\beta}_1\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i y_i \end{cases}$$

16page

(1)에서

$$= \Sigma(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$
$$= \Sigma(y_i - \hat{y}_i)$$
$$= \Sigma e_i = 0$$

(2) 16page

$$= \Sigma x_i(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$
$$= \Sigma x_i e_i = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \hat{\beta}_1 \frac{1}{n}\sum x_i$$

$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

(2)식 $\sum x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$$\sum x_i(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)$$

$$= \boxed{\underbrace{\sum x_i (y_i - \bar{y})}_{㉠}} - \boxed{\underbrace{\sum x_i \hat{\beta}_1 (x_i - \bar{x})}_{㉡} = 0}$$

㉠ $= \sum (x_i - \bar{x} + \bar{x})(y_i - \bar{y})$

$$= \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum \bar{x}(y_i - \bar{y})$$

$$= \boxed{S_{un(xy)}}$$

x편차와
y편차곱의합

$x_i - \bar{x}$: 편차
$\sum(x_i - \bar{x}) = 0$
pf) $\sum x_i - \sum \bar{x}$
$= \sum x_i - n \cdot \bar{x}$
$= \sum x_i - \frac{1}{n}\sum x_i \times n$
$= 0$

㉡ $= \hat{\beta}_1 \sum x_i (x_i - \bar{x})$

$$= \hat{\beta}_1 \left(\sum (x_i - \bar{x} + \bar{x})(x_i - \bar{x})\right)$$

$$= \hat{\beta}_1 \left(\sum (x_i - \bar{x})(x_i - \bar{x}) + \sum \bar{x}(x_i - \bar{x})\right)$$

$$= \hat{\beta}_1 (x_i - \bar{x})^2$$

$$= \boxed{\hat{\beta}_1 S_{(xx)}}$$

$$\Rightarrow \quad S_{(xy)} - \hat{\beta}_1 S_{(xx)} = 0.$$

$$\hat{\beta}_1 = \frac{S_{(xy)}}{S_{(xx)}}$$

Least Square Estimation (LSE)

- 최소제곱추정량(LSE)

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{S_{(xy)}}{S_{(xx)}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

- $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i, \ \bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

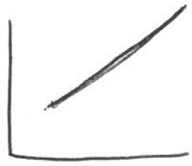
Example - LSE

광고비 판매액 S_{xx} 구할때 S_{xy} 구할때

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	4.00	9.00	-4.00	-9.60	16.00	92.16	38.40
2	8.00	20.00	0.00	1.40	0.00	1.96	0.00
3	9.00	22.00	1.00	3.40	1.00	11.56	3.40
4	8.00	15.00	0.00	-3.60	0.00	12.96	-0.00
5	8.00	17.00	0.00	-1.60	0.00	2.56	-0.00
6	12.00	30.00	4.00	11.40	16.00	129.96	45.60
7	6.00	18.00	-2.00	-0.60	4.00	0.36	1.20
8	10.00	25.00	2.00	6.40	4.00	40.96	12.80
9	6.00	10.00	-2.00	-8.60	4.00	73.96	17.20
10	9.00	20.00	1.00	1.40	1.00	1.96	1.40
sum	80	186	0	0	46	368.4	120

$\bar{x} = 8$ $\bar{y} = 18.6$ $S_{(xx)}$ $S_{(yy)}$ $S_{(xy)}$

Example - LSE



- 최소제곱추정량 (LSE)

$$\hat{\beta}_1 = \frac{120}{46} = 2.6087,$$

$$\hat{\beta}_0 = 18.6 - 2.6087 \times 8 = -2.2269$$

- 추정된 회귀직선: Least square fit

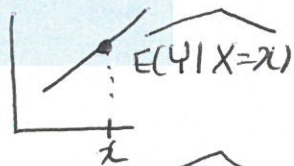
$$\hat{y} = \underbrace{-2.2696}_{\beta_0} + \underbrace{2.6087}_{\beta_1} \cdot x$$

기울기 $(\beta_1) = \frac{\Delta y}{\Delta x}$: x가 1단위 증가할 때 y의 변화량

∴ 광고비가 (x) 10만원 증가하면
총판매액은 2.6087 X(백만) 증가한다.
≒ 260만

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

Properties of fitted regression line



$\hat{y}_i = E(y_i \mid X = x_i)$

*편차: $x_i - \bar{x}$
오차: $\mu - \hat{\mu}$ 실제값과 추정값의 차 (모수에서)
잔차: $y_i - \hat{y}_i$
오차항의 추정량

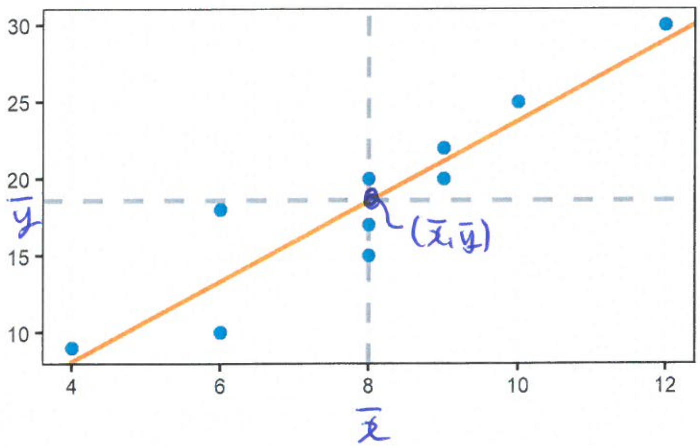
- 잔차 (residual) : $\boxed{e_i = y_i - \hat{y}_i,}$ $i = 1, 2, \ldots, n$

$e_i = \hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
실제값과 추정된 회귀직선의 차이 잔차

(1) 잔차의 합은 0이다. $(\sum_{i=1}^{n} e_i = 0)$ 12page
(2) $\Sigma_{i=1}^{n} e_i^2$ 은 최소값을 갖는다. 10page
(3) 잔차의 x_i에 의한 가중합은 0이다. $(\sum_{i=1}^{n} x_i e_i = 0)$ 12page
(4) 잔차의 $\hat{y}_i$에 의한 가중합은 0이다. $(\sum_{i=1}^{n} \hat{y}_i e_i = 0)$
(5) $(\bar{x}, \bar{y})$는 적합된 회귀직선 위에 있다.

$= \Sigma(\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i$
$= \Sigma \hat{\beta}_0 e_i + \Sigma \hat{\beta}_1 x_i e_i$
$= \underbrace{\hat{\beta}_0 \Sigma e_i}_{(1)에 의해 =0} + \underbrace{\hat{\beta}_1 \Sigma x_i e_i}_{(3)에 의해 =0}$
$= 0$

Example



$x_1, x_2, \cdots, x_n \sim N(\mu, \sigma^2)$ 일때

$\hat{\mu} = \frac{1}{n}\sum x_i$

$\hat{\sigma}^2 = \underbrace{\frac{1}{n-1}}_{자유도}\sum(x_i - \bar{x})^2$

*불편추정량 : θ 모수 → $\hat{\theta}$
(unbiased estimate)

⇒ $E(\hat{\theta}) = \theta$

* $E(\hat{\mu}) = \mu$ (∵정규재생특징)
$E(\hat{\sigma}^2) = \sigma^2$ (카이제곱 〃)

$y = \beta_0 + \beta_1 x + \varepsilon_i$
$\varepsilon_i \sim N(0, \sigma^2)$ iid
추정해야함

Estimation of error variance

$e_1, e_2, \cdots, e_n$

$\hat{\sigma}^2 = \frac{1}{n-2}\sum(e_i - \bar{e})^2$

($\bar{e} = \frac{1}{n}\sum e_i = 0$)

$= \frac{1}{n-2}\sum e_i = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

(제약조건에 의해 자유도 2개가 빠짐)

($\sum e_i = 0$
$\sum x_i e_i = 0$)

■ 오차분산 (σ^2)의 추정:

- 잔차(오차) 제곱합 (residual (or error) sum of squares) :

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

- 평균제곱오차 (mean squared error) : $MSE = \frac{SSE}{n-2}$
- 오차분산의 추정값 : $\hat{\sigma}^2 = MSE$

Decomposition of deviations

- 총편차의 분해
  - $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \ \forall i$
  - 총편차(total deviation) $= y_i - \bar{y}$
  - 추측값의 편차 $= (\hat{y}_i - \bar{\hat{y}}) = (\hat{y}_i - \bar{y})$

  $\Rightarrow$ 총편차 = $\sum$잔차^2 + $\sum$추측값의 편차^2

  부호 중요X. 절대값!

Decomposition of deviations

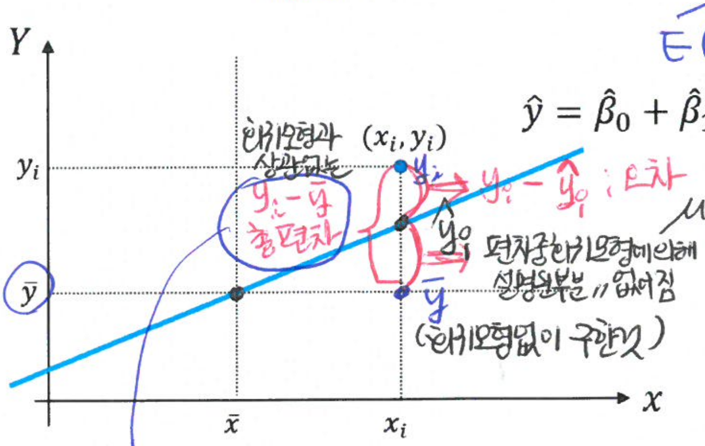
x가 없을때,

$E(y_i) = \mu$

$\hat{\mu} = \bar{y}$

Figure: 편차의 분해

$E(Y|X=x)$



$\mu_{y_i|x_i} = E(y_i | X = x_i)$

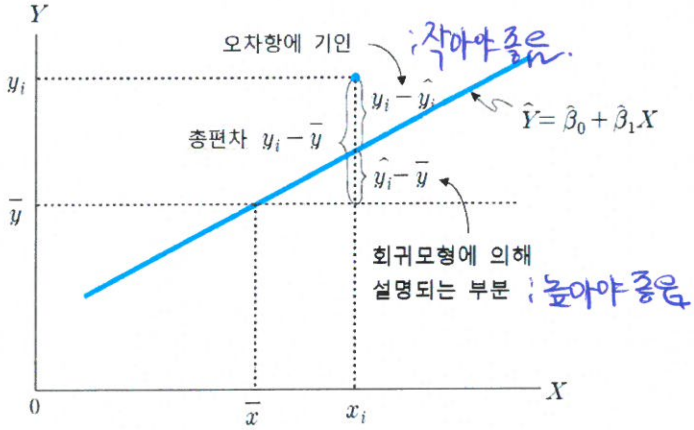
$\hat{\mu}_{y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$

x가 있을때

설명변수와 관계없이 항상 일정

Decomposition of deviations

Figure: 편차의 분해



모형적합이 잘 됐다는 뜻은?

→ 잔차는 작게, 회귀모형에의해 설명되는 부분은 크게

Decomposition of sum of squares

- 제곱합의 분해 : SST = SSE + SSR

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

제곱합의 종류	정의 및 기호
총제곱합 (total sum of squares)	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
잔차제곱합 (residual sum of squares) (에러)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ ↓
회귀제곱합 (regression sum of squares)	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ ↑

편차 $y_i - \bar{y} = (y_i - \hat{y}_i + \hat{y}_i - \bar{y})$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum (y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$= \sum\left((y_i-\hat{y}_i)^2+(\hat{y}_i-\bar{y})^2+2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})\right)$$

$$= \sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2+2\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$\sum e_i(\hat{y}_i-\bar{y})$$

$$= \underbrace{\sum e_i\hat{y}_i}_{\text{by (4)} \ = 0} - \underbrace{\sum \bar{y}\cdot e_i}_{\bar{y}\sum e_i = 0}$$

$$\therefore \boxed{\sum(y_i-\bar{y})^2 = \sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2}$$

SST 총제곱합   SSE 잔차 제곱합   SSR 회귀제곱합

(잔차 $= e_i$)
residual

Coefficient of determination

■ 결정계수 (Coefficient of determination)

- 정의 : $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$
- 의미 : 회귀직선의 기여율

  (총변동 가운데 회귀직선으로 설면되는 변동의 비율)
- 성질
  - $0 \le R^2 \le 1$
  - R^2 값이 1에 가까울수록 회귀에 의한 설명이 잘 됨을 뜻함
  - $R^2 = r^2$ (r : sample correlation)

    (단순선형회귀모형에서만 성립)

R^2은 비교할때 사용 (상대적임)
Model 1 80%. Model2 75%인데

상관분석

$\rho > 0$: 양의상관관계
$\rho < 0$: 음의상관관계
$\rho \simeq 0$: 상관관계가 없다.

- X, Y : random variables

ρ 로우

- 모상관계수 (population coefficient of correlation)

$$\rho_{XY} \quad \rho_{XY} = \frac{\mathrm{Cov(X,Y)}}{\sqrt{\mathrm{Var(X)Var(Y)}}} := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- $(x_1, y_1), \ldots, (x_n \cdot y_n)$: sample
- 표본상관계수 (sample coefficient of correlation)

$$\hat{\rho}_{XY} = \boxed{r_{xy} = \frac{S_{(xy)}}{\sqrt{S_{(xx)} S_{(yy)}}}}$$

- $-1 \le \rho \le 1, \quad -1 \le r_{xy} \le 1$

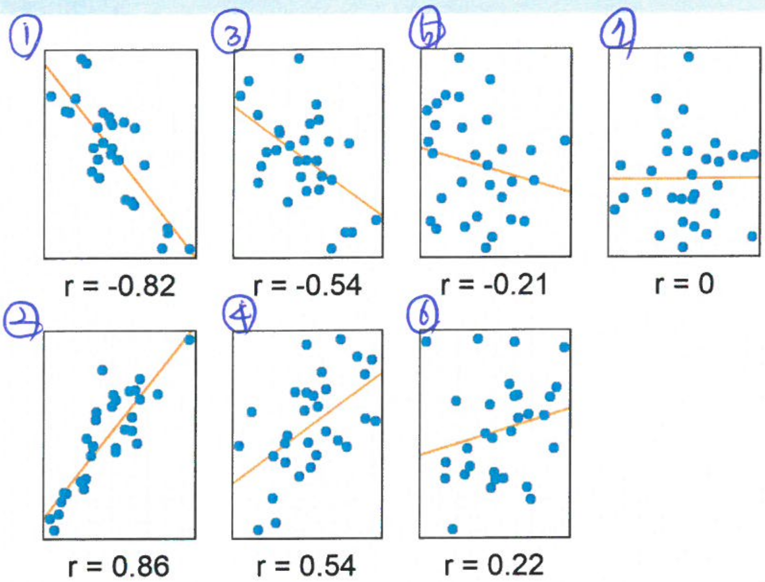
$$S_X^2 = \frac{1}{n-1} \Sigma (x_i - \bar{x})^2 = \frac{1}{n-1} S_{(xx)}$$

$$S_Y^2 = \frac{1}{n-1} S_{(yy)}$$

$$\mathrm{Cov}(X,Y) = \sigma_{XY} = E(X-\mu_X)(Y-\mu_Y)$$

$$\Rightarrow \hat{\sigma}_{XY} = \frac{1}{n-1} \Sigma (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} S_{(xy)}$$

표본상관계수와 산점도



①, ② 상관관계가 강한 case

Example

- 광고료와 총판매액
  - 표본상관계수

$$r_{xy} = \frac{S_{(xy)}}{\sqrt{S_{(xx)}S_{(yy)}}} = \frac{120}{\sqrt{46 \times 368.4}} = 0.92$$

(1에 가까움.
상관관계가 크다.

표본상관계수와 단순선형회귀모형

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{S^2_{(xy)}}{S_{(yy)}S_{(xx)}} = r^2_{xy}$$

■ 단순선형회귀모형에서는

- 표본상관계수와 결정계수

$$\boxed{R^2 = r_{xy}^2}$$

$$SSR : \Sigma(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$
$$= \Sigma(\cancel{\bar{y}} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \cancel{\bar{y}})^2$$
$$= \Sigma \hat{\beta}_1^2 (x_i - \bar{x})^2$$
$$= \hat{\beta}_1^2 \underbrace{\Sigma(x_i - \bar{x})^2}_{S(xx)}$$
$$\left(\hat{\beta}_1 = \frac{S(xy)}{S(xx)}\right)$$
$$= \frac{S^2(xy)}{S(xx)}$$

- 표본상관계수와 회귀직선의 기울기

$\hat{\beta}_1 = \oplus \; r_{xy} \uparrow 1$
$\quad\;\; \ominus \; r_{xy} \downarrow -1$

$$\boxed{r_{xy} = \hat{\beta}_1 \frac{s_x}{s_y}}$$

s_x, s_y : 표본표준편차 (sample standard deviation)

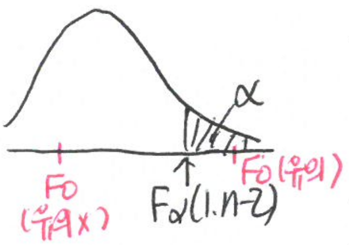
$$s_x = \sqrt{\frac{S_{(xx)}}{n-1}}, \quad s_y = \sqrt{\frac{S_{(yy)}}{n-1}}$$

분산분석

SST = SSR + SSE
↑ 커야좋음

■ 단순회귀직선의 유의성 검정을 위한 분산분석표

설명변수의 갯수



요인	제곱합(SS)	자유도(df)	평균제곱(MS)	F_0	유의확률
회귀	SSR	1	$MSR = \frac{SSR}{1}$	$\frac{MSR \uparrow}{MSE \downarrow}$	$P(F \geq F_0)$
잔차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계	SST	$n-1$			

F0 값이 큰값이된다 = 회귀모형이 잘되었다.
F0↑ : 통계적으로 유의하다.

- $F \sim F(1, n-2)$ 분자 분모 의 자유도
- $F_0 > F(1, n-2; \alpha) \Rightarrow$ 유의수준 α 하에서, 회귀선이 유의함.