제2장 단순선형회귀모형

2.1 회귀분석의 기초개념

1. 회귀(regression)

Francis Galton(1885), "Regression Toward Mediocrity in Hereditary Stature"

- 아버지와 아들의 키에 대한 관련성을 연구하면서 '회귀'란 용어를 처음 사용
- 분석결과, 키가 매우 큰(작은) 부모의 아들은 평균보다는 큰(작은) 키를 가지지만 아버지의 키보다는 작은(큰) 경향이 있었다.

2. 회귀분석이란?

■ 어떤 현상이 변수들의 인과관계에 의해서 나타날 때 그 관계를 수학적으로 설명하기 위해 사용되는 통계적 방법들 중 하나

■ 변수들 간의 관계를 나타내는 타당한 수학적 모형을 이론적 근거나 경험에 바탕하여 설정하고, 변수들의 관측된 값을 이용하여 그 모형을 추정한 다음, 추정한 모형에 의해 변수들 간의 관계를 설명하든지 또는 예측 등의 분석에 응용하게 된다.

3. 회귀분석의 목적

(1) 변수들 간에 성립하는 정확한 회귀모형의 구축

(2) 모형에 포함된 모수들의 추정

(3) 적합된 모형을 이용한 예측

4. 기호: x - 독립변수, 설명변수, 예측변수, 입력변수, y - 종속변수, 반응변수

5. 회귀모형: $y = f(x_1, x_2, \cdots, x_n) + \epsilon$

6. 회귀분석의 순서

(1) 입력변수 x와 반응변수 y 선택

x (입력변수)	y (반응변수)
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$
x_n	y_n

(2) 산점도(scatter plot) 그리기, 기초통계량/요약통계량 계산하기

(3) 회귀모형의 유형 결정

(4) 통계적 추론

(5) 회귀분석 결과 해석 및 응용분야에서의 함의 도출

[예제]

알레르기에 대한 새로운 약품을 개발하는 단계에서 알레르기 증상이 없어지는 약의 지속효과에 영향을 주는 약의 복용량이 어떻게 다른지 알고 싶다. 10명의 환자를 대상으로 각 환자는 규정량의 약을 복용한 후 약의 효과가 사라지면 곧 돌아와 보고하도록 하였다. 10명의 환자에 대한 약의 복용량(x)과 약의 지속효과 기간(y)이 [표]에 주어져 있다. 표를 보면 y가 x에 따라 대체로 증가하는 것으로 보인다.

| 표 10-1 | 10명의 환자에 대한 약의 복용량(x)과 지속효과(y)

약의 복용량 x	약의 지속효과 y
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

| 표 10-2 | 단순회귀에 대한 자료구조

독립변수	반응변수
x_1	y_1
x_2	y_2
x_3	y_3
.	.
.	.
.	.
x_n	y_n

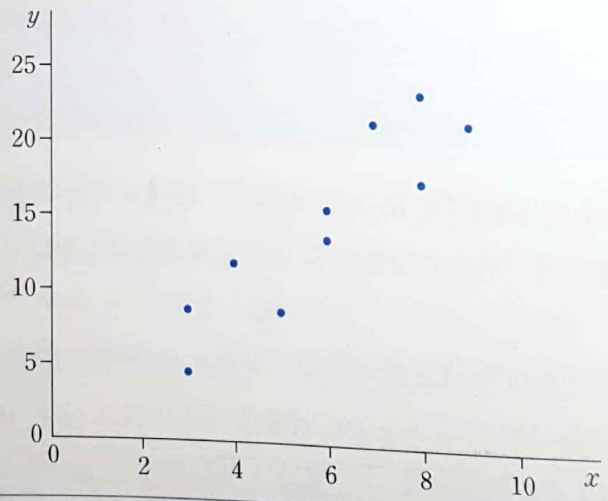


그림 10-1 | [표 10.1]에 주어진 자료의 산점도

2.2 단순선형회귀모형

1. 단순선형회귀에 대한 통계적 모형

반응변수(y)는 $y_i = \beta_0 + \beta_1 x_i + e_i \ \ (i = 1, \cdots, n)$에 의해 입력변수$(x)$와 관련되어지는 확률변수라고 가정한다. 이때

(1) y_i는 설명변수 x가 x_i일 때의 반응치이다.

(2) $e_1, \cdots, e_n$은 실제 직선관계에 부과되는 알 수 없는 오차요소들이다. 이것들은 평균이 0이고 표준편차가 σ인 정규분포를 따르는 확률변수이다.

$$E(y_i) = E(\beta_0 + \beta_1 x_i + e_i) = E(\beta_0 + \beta_1 x_i) + E(e_i) = \beta_0 + \beta_1 x_i$$

$$Var(y_i) = Var(\beta_0 + \beta_1 x_i + e_i) = Var(\beta_0 + \beta_1 x_i) + Var(e_i) = \sigma^2$$

(3) β_0와 β_1은 미지의 계수이다.

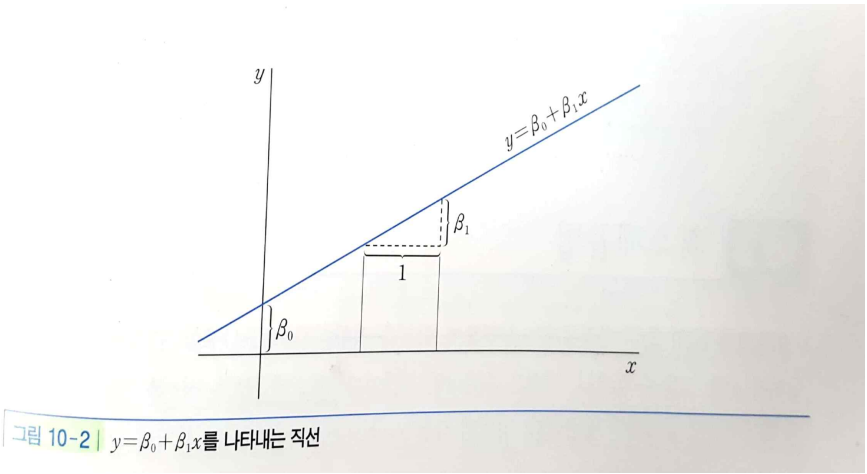


그림 10-2 | $y=\beta_0+\beta_1 x$를 나타내는 직선

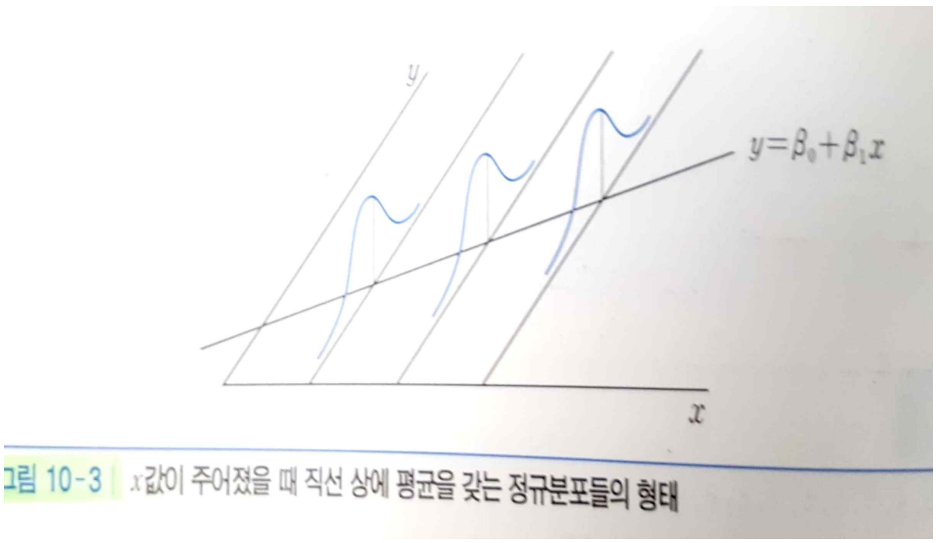
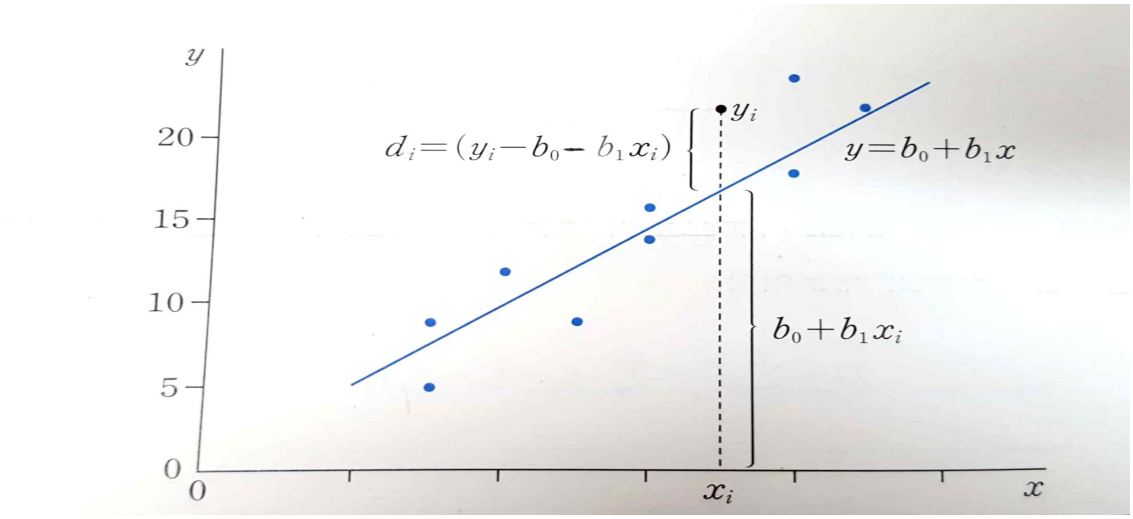


그림 10-3 | x값이 주어졌을 때 직선 상에 평균을 갖는 정규분포들의 형태

2.3 회귀계수의 추정

2.3.1 최소제곱법



1. 최소제곱법의 원리

① $D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$가 최소가 되는 모수값을 결정하는 것

② 최소제곱추정량(least squares estimator: LSE): $\hat{\beta}_0,\ \hat{\beta}_1$

③ 적합된 직선: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

2. 기본적인 기호

① $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

② $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

③ $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$

④ $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i \right)^2}{n}$

⑤ $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i \right)\left(\sum_{i=1}^{n} y_i \right)}{n}$

3. 최소제곱추정량에 대한 공식

(1) β_0의 최소제곱추정량: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

(2) β_1의 최소제곱추정량: $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

$$D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial D}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \rightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \qquad \text{정규방정식}$$

$$\frac{\partial D}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \qquad \text{정규방정식}$$

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i &= (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\hat{\beta}_1 + n\bar{x}\bar{y}\end{aligned}$$

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

4. 잔차: $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \ \ (i = 1, \cdots, n)$

(1) 잔차: 0, 양수, 음수

(2) 잔차의 합은 언제나 0이다.

(3) 잔차제곱합(residual sum of squares)

[오차제곱합(sum of squares due to error)] → SSE

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

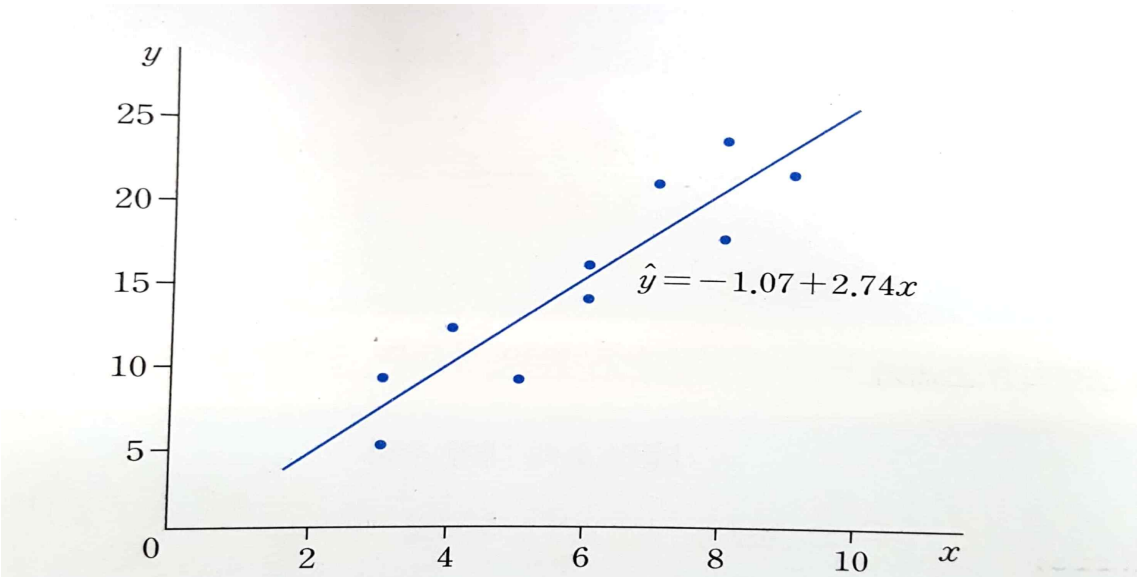
(4) 모형검토에서 잔차의 역할이 중요하다.

5. 분산의 추정: 분산 σ^2은 $s^2 = \dfrac{SSE}{n-2}$에 의해 추정된다.

6. 예제

	x	y	x^2	y^2	xy	$\hat{\beta}_0+\hat{\beta}_1 x$	$\hat{e}$
	3	9	9	81	27	7.15	1.85
	3	5	9	25	15	7.15	-2.15
	4	12	16	144	48	9.89	2.11
	5	9	25	81	45	12.63	-3.63
	6	14	36	196	84	15.37	-1.37
	6	16	36	256	96	15.37	0.63
	7	22	49	484	154	18.11	3.89
	8	18	64	324	144	20.85	-2.85
	8	24	64	576	192	20.85	3.15
	9	22	81	484	198	23.59	-1.59
합계	59	151	389	2,651	1,003		0.04

$\bar{x}=5.9,\ \bar{y}=15.1$ $S_{xx}=389-\frac{59^2}{10}=40.9$ $S_{yy}=2651-\frac{151^2}{10}=370.9$ $S_{xy}=1003-\frac{59\times151}{10}=112.1$	$\hat{\beta}_1=\frac{112.1}{40.9}=2.74$ $\hat{\beta}_0=15.1-2.74\times5.9=-1.07$ $SSE=370.9-\frac{112.1^2}{40.9}=63.6528$ $s^2=\frac{SSE}{n-2}=\frac{63.6528}{8}=7.96$
$\hat{y}=-1.07+2.74x$	



2.3.2 최소제곱추정량의 성질

1. 불편성

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{y}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}} = \sum_{i=1}^{n} w_i y_i$$

여기서, $w_i = \dfrac{x_i-\bar{x}}{S_{xx}}$

→ $\hat{\beta}_1$은 y_i들의 선형결합

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n}\frac{x_i-\bar{x}}{S_{xx}} = \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2 = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{S_{xx}}$$

$$\Rightarrow\ E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} w_i y_i\right) = \sum_{i=1}^{n} w_i E(y_i) = \sum_{i=1}^{n} w_i(\beta_0+\beta_1 x_i) = \beta_0\sum_{i=1}^{n} w_i + \beta_1\sum_{i=1}^{n} w_i x_i = \beta_1$$

여기서, $\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n}\left(\dfrac{x_i-\bar{x}}{S_{xx}}\right)x_i = \dfrac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})x_i = \dfrac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = 1$

$\therefore\ E(\hat{\beta}_1) = \beta_1$: $\hat{\beta}_1$은 β_1의불편 추정량

$$E(\hat{\beta}_0) = E(\bar{y}-\hat{\beta}_1\bar{x}) = E(\bar{y}) - \bar{x}E(\hat{\beta}_1) = \beta_0+\beta_1\bar{x}-\beta_1\bar{x} = \beta_0$$

여기서, $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{n}\sum_{i=1}^{n}(\beta_0+\beta_1 x_i+\epsilon_i) = \beta_0+\beta_1\bar{x}+\bar{\epsilon}$

$$E(\bar{y}) = E(\beta_0+\beta_1\bar{x}+\bar{\epsilon}) = \beta_0+\beta_1\bar{x}+E(\bar{\epsilon}) = \beta_0+\beta_1\bar{x}$$

$\therefore\ E(\hat{\beta}_0) = \beta_0$: $\hat{\beta}_0$은 β_0의불편 추정량