제10장 회귀분석

10.1 서론

(1) 회귀분석이란?
변수들 간의 관련성에 관한 것으로, 예측을 위한 모형의 구축 및 여러 가지 추론을 하기 위한 분야

(2) 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.

(3) 기호: x - 독립변수, 설명변수, 예측변수, 입력변수, y - 종속변수, 반응변수

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

(5) 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 10.4]
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 10.1]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

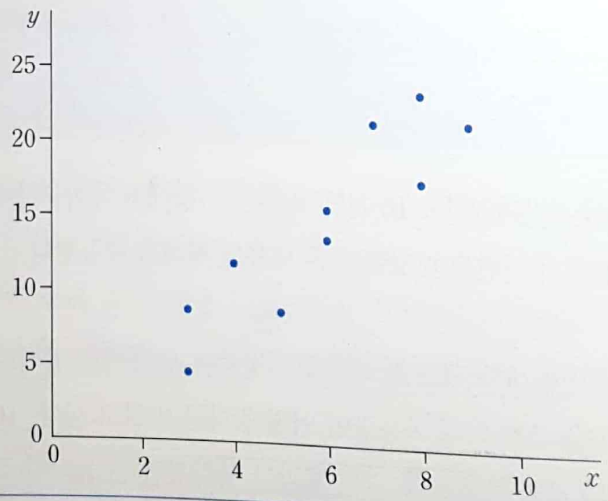


그림 10-1 | [표 10.1]에 주어진 자료의 산점도

10.2 직선회귀모형

(1) 직선회귀에 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.

③ β_0와 β_1은 미지의 계수이다.

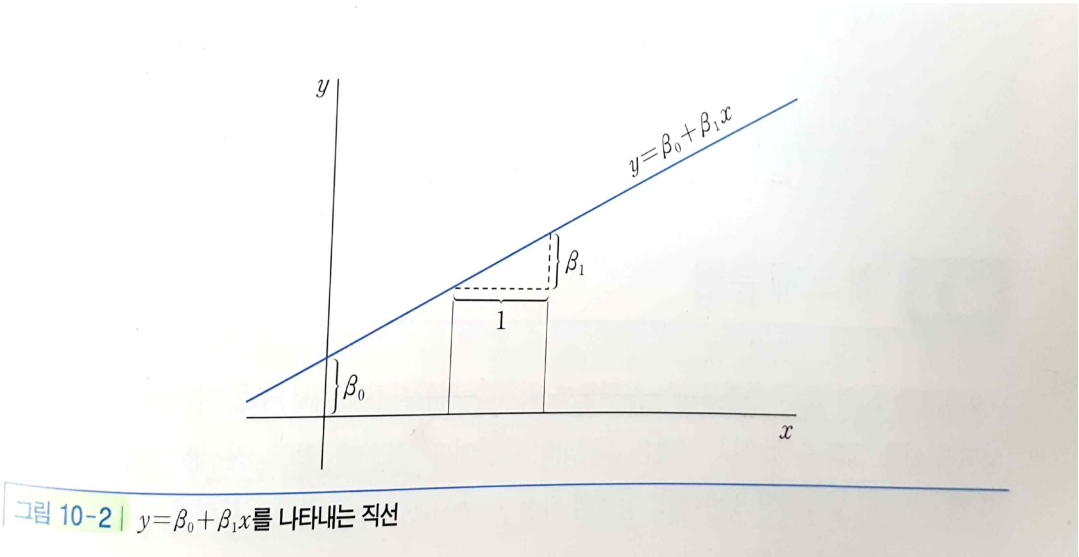


그림 10-2 | $y=\beta_0+\beta_1 x$를 나타내는 직선

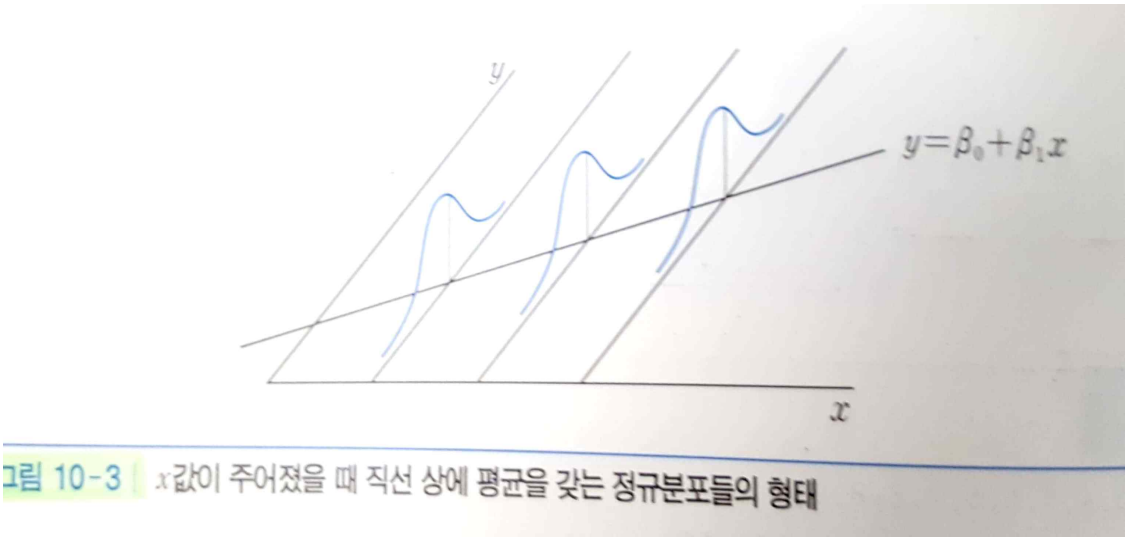
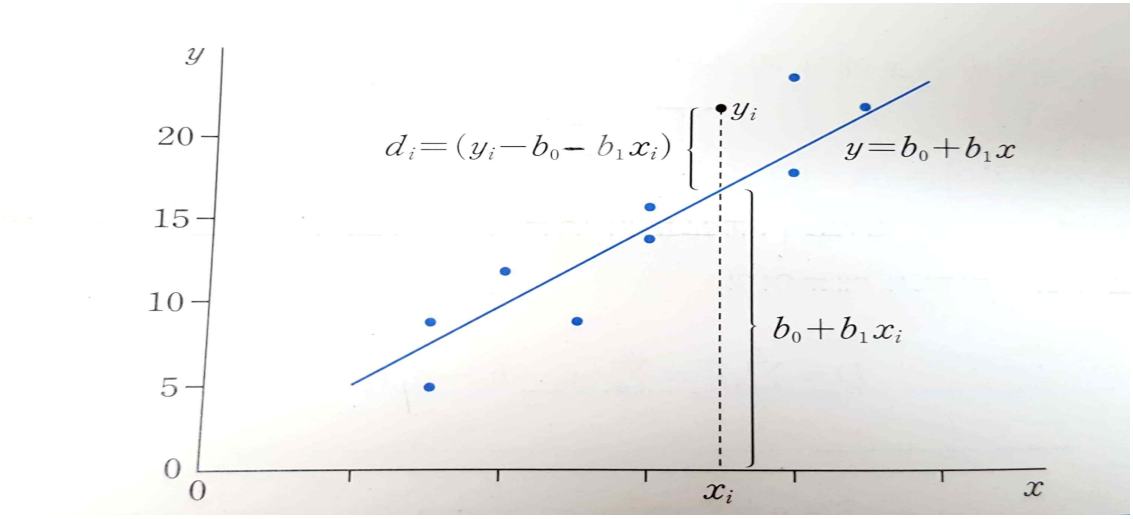


그림 10-3 | x값이 주어졌을 때 직선 상에 평균을 갖는 정규분포들의 형태

10.3 최소제곱법



(1) 최소제곱법의 원리

① $D=\sum_{i=1}^{n}d_i^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$가 최소가 되는 모수값을 결정하는 것

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

③ 적합된 직선: $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$

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

(3) 최소제곱추정량에 대한 공식

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

② β_1의 최소제곱추정량: $\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$

(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

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

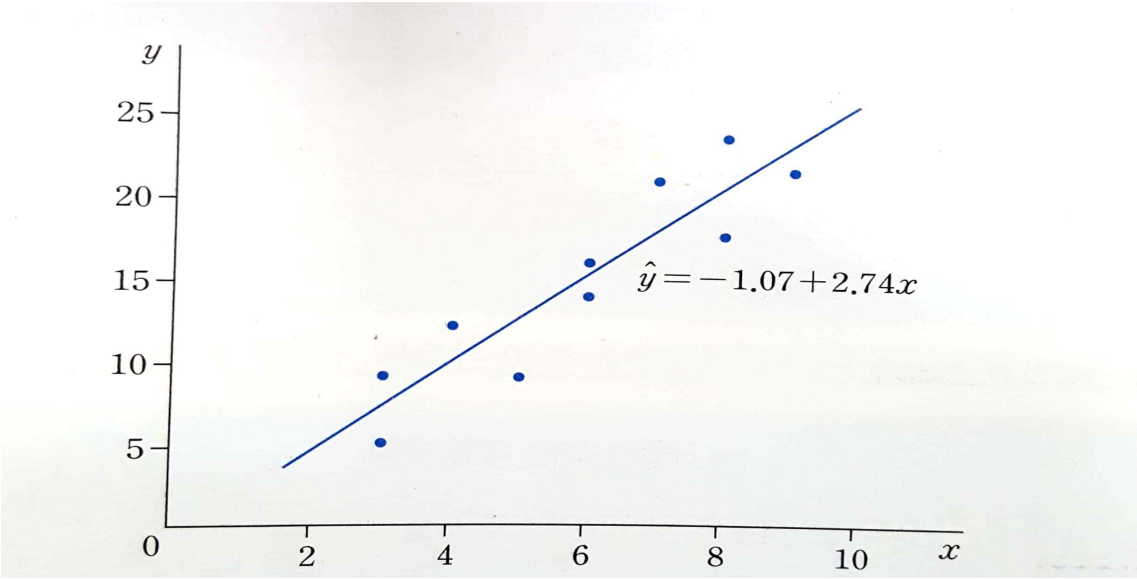
④ 모형검토에서 잔차의 역할이 중요하다.

(5) 분산의 추정: 분산 σ^2은 $s^2 = \frac{SSE}{n-2}$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

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



10.4 최소제곱추정량의 표본분포

(1) 질문

[예제 10.4] 추정식 $\hat{y} = -1.07 + 2.74x$

< 추정치와 관련된 두 가지 질문 >

① $\hat{\beta}_1$에 대한 값 2.74에 비추어 보아 실제 회귀식의 기울기 β_1은 얼마인가?

② 주어진 복용량 $x^* = 4.5mg$에 해당하는 11.26일의 추정된 약의 지속기간은 얼마나 많은 불확실성을 내포하고 있는가?

(2) 상기 질문들에 답하기 위해서는 '최소제곱추정량의 표본분포'를 알아야 하며, 이때 $t-$분포가 이용된다.

(3) 최소제곱추정량의 표준편차(표준오차)

$$S.E.(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$$

$$S.E.(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

추정된 표준오차를 계산하기 위해서는 σ 대신 $s = \sqrt{\frac{SSE}{n-2}}$ 를 사용한다.

(4) 기울기 β_1에 대한 추정

검정통계량 $t=\dfrac{\hat{\beta}_1-\beta_1}{s/\sqrt{S_{xx}}}$는 자유도 $(n-2)$인 t-분포를 따른다.

(5) 절편 β_0에 대한 추정

검정통계량 $t=\dfrac{\hat{\beta}_0-\beta_0}{s\sqrt{\dfrac{1}{n}+\dfrac{\bar{x}^2}{S_{xx}}}}$는 자유도 $(n-2)$인 t-분포를 따른다.

(6) $x=x^*$에서의 반응치의 기댓값: $\beta_0+\beta_1 x^*$

이것은 $\hat{\beta}_0+\hat{\beta}_1 x^*$에 의해 추정되고, 추정된 표준오차는 $s\sqrt{\dfrac{1}{n}+\dfrac{(x^*-\bar{x})^2}{S_{xx}}}$ 이다.

$\hat{\beta}_0+\hat{\beta}_1 x^*$에 대한 추정의 검정통계량

$T=\dfrac{(\hat{\beta}_0+\hat{\beta}_1 x^*)-(\beta_0+\beta_1 x^*)}{s\sqrt{\dfrac{1}{n}+\dfrac{(x^*-\bar{x})^2}{S_{xx}}}}$는 자유도 $(n-2)$인 t-분포를 따른다.

10.5 중요한 추정문제

10.5.1 기울기 β_1에 관련된 추정

(1) 관심문제: 반응변수의 기대치가 입력변수 x의 크기에 따라 변하는가?

(2) 가설설정: $H_0 : \beta_1 = 0$

(3) 검정통계량: $T=\dfrac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$, 자유도=$n-2$

[예제 10.5]

[표 10.1]에서 주어진 자료는 약의 지속효과가 약의 복용량에 따라 증가함을 보여주는가?

【풀이】

① 가설설정: $H_0 : \beta_1 = 0,\ H_1 : \beta_1 > 0$ (단측검정)

② 주어진 정보와 검정통계량

$\hat{\beta}_1 = 2.74,\ s^2 = \dfrac{SSE}{n-2} = \dfrac{63.6528}{8} = 7.9566,\ s = 2.8207$

$S.E.(\hat{\beta}_1) = \dfrac{s}{\sqrt{S_{xx}}} = \dfrac{2.8207}{\sqrt{40.90}} = 0.441$

검정통계량: $t = \dfrac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} = \dfrac{2.74}{0.441} = 6.213$

③ 기각치 및 기각역

기각치: $t_{0.05}(8) = 1.860$, 기각역: $R: t \geq 1.860$

④ 결론: H_0 기각, H_1이 맞다고 말할 수 있다.

→ 약의 복용량이 증가하면 약의 지속효과를 증가시키는 경향이 있다.

(4) $H_0 : \beta_1 = \beta_{10}$의 가설검정

검정통계량 $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s/\sqrt{S_{xx}}}$, 자유도=$n-2$

H_1의 형태와 상응하는 기각역

H_1	기각역
$H_1 : \beta_1 > \beta_{10}$	$R: t \geq t_\alpha$
$H_1 : \beta_1 < \beta_{10}$	$R: t \leq -t_\alpha$
$H_1 : \beta_1 \neq \beta_{10}$	$R: \lvert t \rvert \geq t_{\alpha/2}$

(5) β_1에 대한 $100(1-\alpha)\%$ 신뢰구간: $\hat{\beta}_1 \pm t_{\alpha/2}\dfrac{s}{\sqrt{S_{xx}}}$

[예제 10.6]

[표 10.1]자료를 참고한 회귀직선의 기울기에 대한 95% 신뢰구간을 구하라.

【풀이】

$\hat{\beta}_1 = 2.74,\ S.E.(\hat{\beta}_1) = \dfrac{s}{\sqrt{S_{xx}}} = \dfrac{2.8207}{\sqrt{40.90}} = 0.441$

β_1에 대한 95% 신뢰구간은 $2.74 \pm 2.306 \times 0.441 = 2.74 \pm 1.02$ 즉 $(1.72, 3.76)$

10.5.2 절편 β_0에 대한 추정

(1) β_0에 대한 $100(1-\alpha)\%$ 신뢰구간: $\hat{\beta}_0 \pm t_{\alpha/2} \times s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$

[예제]

[표 10.1]자료를 참고한 회귀직선의 절편에 대한 95% 신뢰구간을 구하라.

【풀이】

$\hat{\beta}_0 = -1.07,\ \bar{x} = 5.9,\ S_{xx} = 40.9,\ s = 2.8207$

β_0에 대한 95% 신뢰구간은 $-1.07 \pm 2.306 \times 2.8207\sqrt{\frac{1}{10}+\frac{5.9^2}{40.9}} = -1.07 \pm 6.34$

즉 $(-7.41, 5.27)$

$\Rightarrow$ β_0는 입력변수 x에 대한 값 0에 대응하는 평균반응을 나타낸다. [예제 10.4]의 약품개발 문제에서 모수 β_0는 실험에 포함된 x의 범위가 3에서 9 사이이기 때문에 실제로 거의 관심이 없다. 직선을 $x=0$으로 해석하는 것은 비현실적이다. 사실 추정치 $\hat{\beta}_0 = -1.07$은 음수이므로 약의 지속기간이 될 수 없다.

10.5.3 고정된 값 x에 대한 평균반응의 예측

(1) 기대반응 $\beta_0 + \beta_1 x^*$에 대한 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} \times s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$$

(2) $H_0 : \beta_0 + \beta_1 x^* = u_0$에 대한 $100(1-\alpha)\%$ 신뢰구간

검정통계량 $t = \dfrac{(\hat{\beta}_0 + \hat{\beta}_1 x^*) - u_0}{s\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}}$는 자유도 $(n-2)$인 t-분포를 따른다.

H_1의 형태와 상응하는 기각역

H_1	기각역
$H_1 : \beta_0 + \beta_1 x^* > u_0$	$R: t \geq t_\alpha$
$H_1 : \beta_0 + \beta_1 x^* < u_0$	$R: t \leq -t_\alpha$
$H_1 : \beta_0 + \beta_1 x^* \neq u_0$	$R: \lvert t \rvert \geq t_{\alpha/2}$

[예제 10.7]

[표 10.1]에 주어진 자료와 [표 10.3]에 주어진 회귀분석에 대한 계산을 다시 고려해 보자. 복용량이 (a) $x^* = 6$일 때, (b) $x^* = 9.5$일 때 약의 지속기간에 대한 신뢰구간을 구하라.

【(a) 풀이】

적합된 회귀선: $\hat{y} = -1.07 + 2.74x$

약의 복용량 $x^* = 6$에 해당하는 약의 지속기간: $\hat{y} = -1.07 + 2.74x^* = -1.07 + (2.74 \times 6) = 15.37$

$$S.E.(\hat{\beta}_0 + \hat{\beta}_1 x^*) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = s\sqrt{\frac{1}{10} + \frac{(6-5.9)^2}{40.9}} = 2.8207 \times 0.3166 = 0.893$$

약의 복용량 $x^* = 6$에서의 약의 평균지속효과에 대한 95% 신뢰구간

$$: (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{0.025} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= 15.37 \pm (2.306 \times 0.893) = 15.37 \pm 2.06$$

즉 $(13.31, 17.43)$

【(b) 풀이】

약의 복용량 $x^* = 9.5$에 해당하는 약의 지속기간: $\hat{y} = -1.07 + 2.74x^* = -1.07 + (2.74 \times 9.5) = 24.96$

$$S.E.(\hat{\beta}_0 + \hat{\beta}_1 x^*) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = s\sqrt{\frac{1}{10} + \frac{(9.5-5.9)^2}{40.9}} = 2.8207 \times 0.6457 = 1.821$$

약의 복용량 $x^* = 9.5$에서의 약의 평균지속효과에 대한 95% 신뢰구간

$$: (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{0.025} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= 24.96 \pm (2.306 \times 1.821) = 24.96 \pm 4.2$$

즉 $(20.76, 29.16)$

10.5.4 고정된 값 x에 대한 하나의 반응의 예측

x^*가 주어졌을 때 하나의 관측치 y를 예측할 때 추정된 표준오차는

$$s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

※ 하나의 관측치는 모집단 분포의 평균(기댓값)보다 더 불확실하기 때문에 여기서 추정된 표준오차는 더 커진다.

[예제 10.8]

한 번 더 [표 10.1]에 주어진 약품 실험자료를 고려하자. 새로운 실험은 복용량 $x^* = 6.5mg$을 가진 한 명의 환자에게 시행된다. 약의 지속기간을 예측하고 약의 지속기간에 대한 95% 신뢰구간을 구하라.

【풀이】

예측된 약의 지속기간: $\hat{y} = -1.07 + 2.74x^* = -1.07 + (2.74 \times 6.5) = 16.74$

약의 복용량 $x^* = 6.5$에서의 약의 평균지속효과에 대한 95% 신뢰구간

$$: (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} \times s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{0.025} \times s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$= 16.74 \pm 6.85$

즉 $(9.89, 23.59)$

[예제 10.9]

일하는 데 있어 숙련도를 어떻게 측정할 수 있는지의 연구에서 복잡한 조립일은 훈련의 양에 의해 영향을 받는데, 15명의 새로운 신참자에게 3시간에서 12시간까지 다양하게 훈련을 시켰다. 훈련 후에 일을 수행하는 시간이 기록된다. x=훈련의 지속기간(시간), y=일을 하는 시간(분)이라고 할 때, 다음의 요약통계량을 얻었다고 하자.

$$\bar{x} = 7.2,\ \bar{y} = 45.6,\ S_{xx} = 33.6,\ S_{yy} = 160.2,\ S_{xy} = -57.2$$

(1) 최소제곱법에 의해 적합된 직선의 방정식을 구하라.

【풀이】

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-57.2}{33.6} = -1.702$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 45.6 - (-1.702) \times 7.2 = 57.85$$

적합된 직선의 방정식: $\hat{y} = 57.85 - 1.702x$

(2) 작업시간은 훈련시간이 늘어남에 따라 감소하는가?

【풀이】

① 가설설정: $H_0 : \beta_1 = 0,\ H_1 : \beta_1 < 0$ (단측검정)

② 주어진 정보와 검정통계량

$\hat{\beta}_1 = -1.702,\ S^2 = \frac{SSE}{n-2} = \frac{62.824}{13} = 4.8326,\ s = 2.198$

여기서 $SSE = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 160.2 - \frac{(-57.2)^2}{33.6} = 62.824$

$S.E.(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}} = \frac{2.198}{\sqrt{33.6}} = 0.379$

검정통계량: $T = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} = \frac{-1.702}{0.379} = -4.49$

③ 기각치 및 기각역

기각치: $t_{0.05}(13) = -1.771$, 기각역: $R: t \le -1.771$

④ 결론: H_0 기각, H_1이 맞다고 말할 수 있다.

→ 작업시간은 훈련시간이 늘어남에 따라 감소하는 경향이 있다.

(3) 9시간의 훈련에 대한 평균 작업시간을 추정하고, 95% 신뢰구간을 구하라.

【풀이】

훈련의 지속기간(시간) $x^* = 9$에 해당하는 기대 작업시간

: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 57.85 + [(-1.702) \times 9] = 42.53$

$S.E.(\hat{\beta}_0 + \hat{\beta}_1 x^*) = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 2.198\sqrt{\frac{1}{15} + \frac{(9-7.2)^2}{33.6}} = 0.888$

9시간의 훈련에 대한 평균 작업시간에 대한 95% 신뢰구간

$: (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{0.025} \times s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$

$= 42.53 \pm (2.16 \times 0.888) = 42.53 \pm 1.92$

즉 $(40.6,\ 44.5)$

(4) $x = 35$시간에 대한 예측치 y를 구하고 결과를 논하라.

【풀이】

$x = 35$(시간)는 3에서 12의 실험범위를 많이 벗어나므로, 적합된 회귀직선을 사용하여 $x = 35$에서의 y를 예측하는 것은 타당하지 않다.

10.6 직선모형의 적합도 및 모형검토

10.6.1 직선모형의 적합도

(1) 관심문제: 직선모형의 타당성을 위해 우리는 반응변수에서의 얼마나 많은 변동이 적합된 회귀선에 의해 설명되는지 실험한다.

(2) 관측치 y_i의 분해

$$y_i = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) + (y_i - \hat{y}_i) = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) + \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$$

① 관측된 y
② 선형관계에 의해 설명되는 부분
③ 선형관계로부터의 잔차 또는 편차

(3) 변동의 분해

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + SSE$$

① y의 총변동
② 선형관계에 의해 설명되는 변동
③ 잔차 또는 설명 안되는 변동

잔차제곱합: $SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

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

y의 전체변동와 SSE와의 차이: $S_{yy} - SSE = S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{S_{xy}^2}{S_{xx}}$

(4) 선형관계의 정도: 직선모형을 어떻게 잘 적합시키는가에 대한 지표 [결정계수]

$$r^2 = \frac{\text{선형회귀에 의한 } y \text{ 변동}}{\text{전체 } y \text{ 변동}} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

(5) 선형관계의 정도는 표본상관계수 r의 제곱인 $r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$에 의해 측정

[예제 10.10]

[표 10.1]에서 약품의 자료를 생각하자.

[표 10.3]에서 주어진 계산으로부터 $S_{xx}=40.9,\ S_{yy}=370.9,\ S_{xy}=112.1$

적합된 회귀식: $\hat{y}=-1.07+2.74x$ 을 얻었다.

y의 변동 중 얼마만큼이 선형회귀모형에 의해 설명되는가?

【풀이】

$$r^2=\frac{S_{xy}^2}{S_{xx}S_{yy}}=\frac{112.1^2}{40.9\times 370.9}=0.83$$

10.6.2 직선모형의 검토

(1) 문제제기

회귀분석은 몇 가지 가설검정이나 모수의 신뢰구간을 구하는 것으로 충분하지 않다. 주어진 자료가 모형을 공식화하는 과정에서 만들어진 가정을 따르지 않는다면 심각하게 잘못된 결론을 얻을 수 있다. 따라서 어떤 모순을 가지고 있는지 자료에 대한 면밀한 검토가 필수적이다.

(2) 직선모형을 공식화하는데 포함된 가정들

① 선형적인 관계
② 오차항들의 독립
③ 등분산
④ 정규분포

10.7 비선형관계와 선형변환

[예제 10.11]

자동차의 브레이크를 작동시켰을 때 정지할 때까지의 거리를 알아보기 위해 10대의 차들을 특정한 속도로 달리게 하여 정지할 때까지의 거리를 측정하였다. 자료들이 직선에 가까운 형태로 변환될 수 있을까?

【풀이】

[표 10-4] 속력과 정지거리에 대한 자료: 처음속도(x), 정지거리(y)

[표 10-5] 속력과 정지거리의 제곱근에 관한 자료: $x, y'=\sqrt{y}$

[그림 10-9] $\sqrt{y}$로 변환된 자료

| 표 10-4 | 속력과 정지거리에 대한 자료

처음 속도 x	20	20	30	30	30	40	40	50	50	60
정지 거리 y	16.3	26.7	39.2	63.5	51.3	98.4	65.7	104.1	155.6	217.2

| 표 10-5 | 속력과 정지거리의 제곱근에 관한 자료

x	20	20	30	30	30	40	40	50	50	60
$y'=\sqrt{y}$	4.037	5.167	6.261	7.969	7.162	9.920	8.106	10.203	12.471	14.738

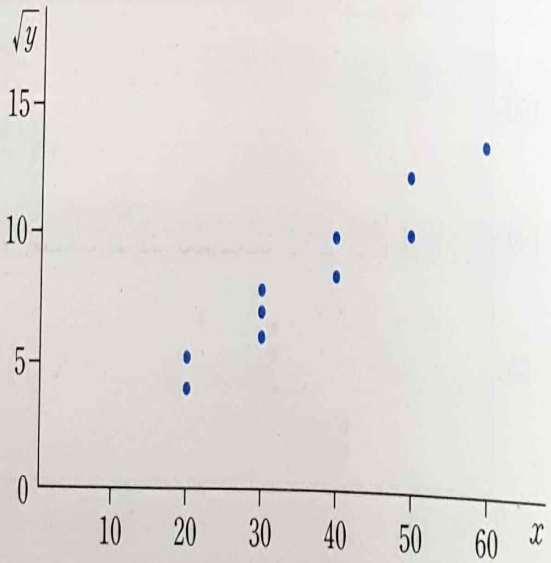


그림 10-9 | $\sqrt{y}$로 변환된 자료([표 10.5]로부터)

$\bar{x}=37,\ \bar{y}'=8.604,\ S_{xx}=1610,\ S_{y'y'}=97.773,\ S_{xy'}=381.621$

$\hat{\beta}_0=-0.671,\ \hat{\beta}_1=0.237$

적합된 직선 방정식: $\hat{y}'=-0.167+0.237x$

직선모형에 의해 설명된 y'의 $r^2=\dfrac{(381.621)^2}{(1610)(97.773)}=0.925$이다.

[표 10-6] 몇 개의 비선형모형과 그들의 선형변환

비선형모형	**변환**	**변환된 모형** $y' = \beta_0 + \beta_1 x'$
$y = ae^{bx}$	$y' = \log_e y \quad x' = x$	$\beta_0 = \log_e a \quad \beta_1 = b$
$y = ax^b$	$y' = \log y \quad x' = \log x$	$\beta_0 = \log a \quad \beta_1 = b$
$y = \frac{1}{a+bx}$	$y' = \frac{1}{y} \quad x' = x$	$\beta_0 = a \quad \beta_1 = b$
$y = a + b\sqrt{x}$	$y' = y \quad x' = \sqrt{x}$	$\beta_0 = a \quad \beta_1 = b$

10.8 다중선형회귀

[표 10-7] 두 개의 설명변수를 가진 다중회귀에 관한 자료구조

실험번호	입력변수		반응변수
	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	y_i
⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	y_n

(1) 다중회귀모형(multiple regression)

① 다중회귀모형

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad (i = 1, \cdots, n)$$

여기서 x_{i1}과 x_{i2}는 i번째 입력변수들의 값, y_i는 그에 대응되는 반응변수값

오차항 e_i는 독립이고 평균이 0, 분산이 σ^2인 정규분포를 따르는 변수

회귀모수 β_0, β_1, β_2와 σ^2은 미지

② 최소제곱법을 이용한 모수추정

▷ $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$을 최소화시켜 얻어짐

▷ 최소제곱추정치 $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$는 다음 방정식의 근이다.

$$\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} = S_{1y}$$

$$\hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} = S_{2y}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

[예제 10.12]

혈압연구에 대해 관심을 가지고 있다. 대략 같은 신장의 남자로부터 혈압 y와 몸무게 x_1, 나이 x_2의 연관성을 보자. 13명으로부터 몸무게, 나이 및 혈압을 측정하였으며 [표 10-8]에 실려 있다.

| 표 10-8 | x_1=몸무게, x_2=나이, y=혈압

x_1	x_2	y
68.9	50	120
83.0	20	141
77.6	20	124
74.8	30	126
71.7	30	117
73.0	50	129
67.6	60	123
71.7	50	125
77.1	40	132
69.4	55	123
74.4	40	132
86.2	40	155
83.9	20	147

[표 10-8] x_1=몸무게, x_2=나이, y=혈압

① 다중회귀모형: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad (i = 1, \cdots, 13)$

② 적합된 회귀직선식: $\hat{y} = -65.02 + 2.3736 x_1 + 0.42542 x_2$ (해석?)

③ 통계적 추론

▷ 모수추정

$\hat{\beta}_0 = -65.02$ $\hat{\beta}_1 = 2.3736$ $\hat{\beta}_2 = 0.42542$

$S.E.(\hat{\beta}_0) = 15.07$ $S.E.(\hat{\beta}_1) = 0.1713$ $S.E.(\hat{\beta}_2) = 0.07377$

표준편차 σ의 추정치: $s = 2.53$

자유도 $= n -$(입력변수의 수)$- 1 = 13 - 2 - 1 = 10$

β_1에 대한 95% 신뢰구간: $2.3736 \pm 2.228 \times 0.1713 = 2.3736 \pm 0.3817$

$(1.9919, 2.7553)$

▷ 평균혈압이 나이에 따라 의미있게 증가하는지 아닌지?

- 가설설정: $H_0 : \beta_2 = 0$ $H_1 : \beta_2 > 0$ (단측검정)
- 검정통계량 $t = 5.77$
- 기각치: $t_{0.01}(10) = 2.764$, 기각역: $R: t \geq 2.764$
- 결론: 귀무가설 기각, 대립가설이 맞다고 말할 수 있다.
  평균혈압이 나이에 따라 통계적으로 의미있게 증가한다고 할 수 있다.

④ 선형모형의 적합성

분산분석표에서 총변동 $\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1486.77$이 두 개의 성분으로 분해된다.

$1486.77 = 1422.80 + 63.97$

▷ y의 총변동

▷ x_1과 x_2에 의해 설명된 변동 따라서 $R^2 = \frac{1422.80}{1486.77} = 0.957$

▷ 잔차 또는 설명 안되는 변동

(2) 다항회귀(polynomial regression)

① 다항회귀모형: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \quad (i = 1, \cdots, n)$

② 다항회귀의 차수: 다항회귀모형에서 발생한 x의 가장 높은 차수

(3) 일반선형모형

① 일반선형모형: $y_i = 1\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i$

② 행렬 표현: $y = X\beta + e$ 여기서 X는 디자인행렬(design matrix)

$$
\begin{aligned}
y_1 &= \beta_0 1 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + e_1 \\
y_2 &= \beta_0 1 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + e_2 \\
&\vdots \\
y_n &= \beta_0 1 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + e_n
\end{aligned}
$$

$$
y_{(n\times 1)} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X_{(n\times(p+1))} = \begin{bmatrix} 1 & x_{11} & x_{12}\cdots x_{1p} \\ 1 & x_{21} & x_{22}\cdots x_{2p} \\ \vdots & & \vdots\ \vdots\ \vdots \\ 1 & x_{n1} & x_{n2}\cdots x_{np} \end{bmatrix},\ \beta_{((p+1)\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},\ e_{(n\times 1)} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}
$$

$$
y = X\beta + e
$$

10.9 모형검토를 위한 잔차분석

(1) 모형검토를 위한 일반적인 방법

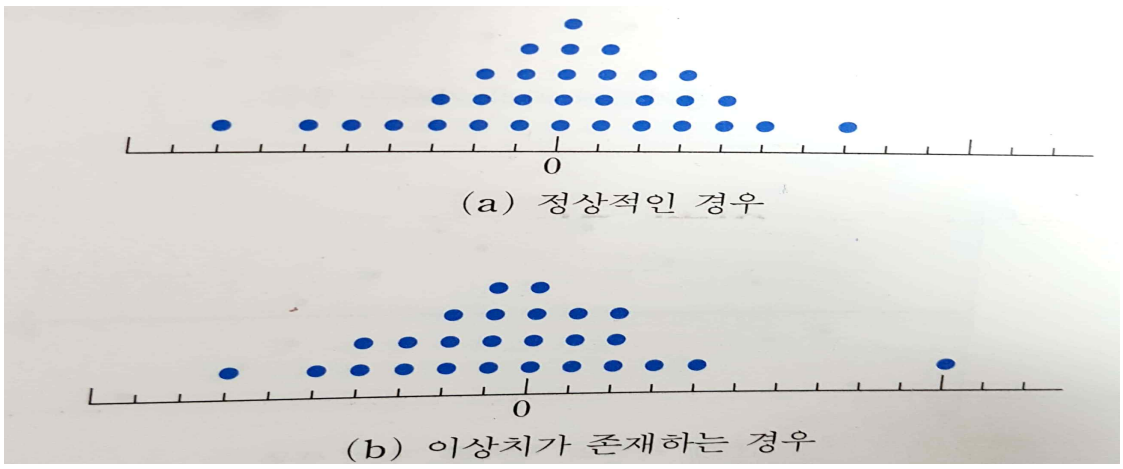
회귀분석은 최소제곱추정량, 신뢰구간, 다양한 가설검정을 구하는 것만은 아니다. 대부분의 경우 특정한 어떤 모형이 맞다는 것을 확신할 수 없다. 따라서 우리는 다음의 단계들을 고려해 보아야 한다.

① 적절한 모형을 고려해 본다.

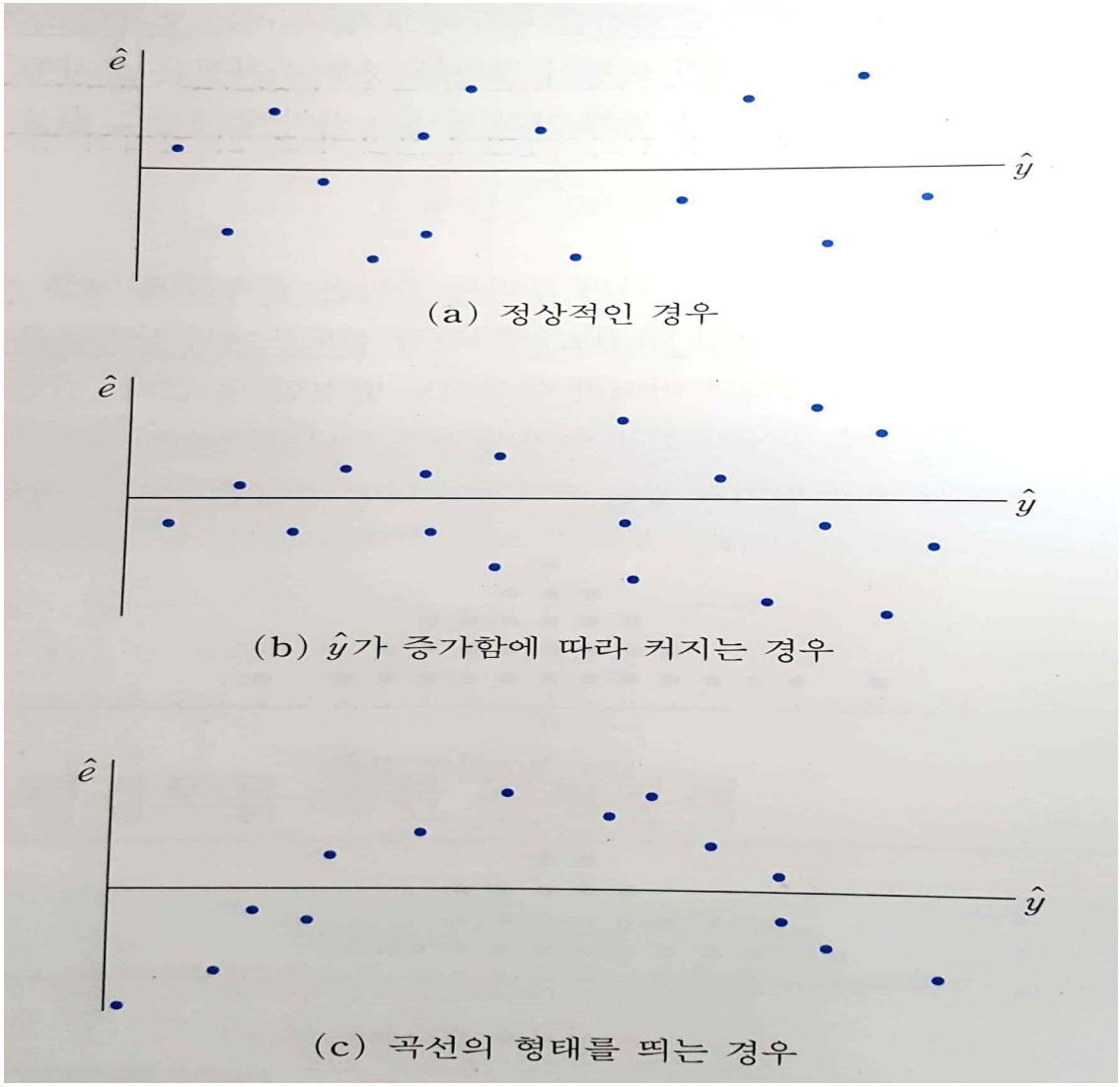
② 최소제곱추정량을 얻고 잔차를 계산한다.

③ 전차를 검토함으로써 모형을 다시 검토한다.

(2) 잔차의 히스토그램 또는 점도표



(4) 잔차와 시간순서의 그림 ~ 독립성 가정의 모순에 대한 검토



(3) 잔차와 예측값의 그림