Seoul National University

제 8장. 상관분석과 회귀분석

서론

- 두 변수 사이의 관계가 관심의 대상이 되는 경우

- **상관분석 (correlation analysis)**
  : 두 변수 사이의 관계 유무 또는 관계의 강도에 대한 통계적 분석 방법
  : 표본 상관계수를 이용한 모상관계수에 대한 추론

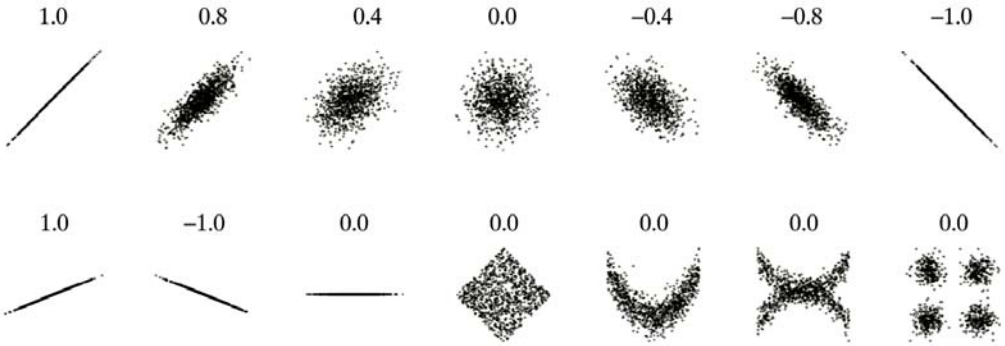
- **회귀분석 (regression analysis)**
  : 두 변수 사이의 함수관계에 대한 통계적 분석
  : 두 변수 사이의 관계식을 파악하여 한 변수의 값으로부터 다른 변수의 값에 대한 예측이 가능하게 됨
  - 선형 회귀분석 (linear regression)
  - 단순 회귀분석 (simple regression)
  - 중회귀분석 (multiple regression)

상관 분석

- 상관계수 (correlation coefficient)
  : 두 변수 사이의 직선관계의 정도를 나타내는 값

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 자료들이 어떤 직선 주위에 밀집되어 나타날수록 -1 또는 1에 가깝게 주어짐
  : (그림) 다양한 산점도의 모양과 표본 상관계수



- 표본상관계수의 값이 0에 가까운 것은 두 변수 사이에 **직선관계**가 약한것을 의미함

상관계수의 검정

- 모집단의 분포가 이변량 정규모집단의 경우
- $H_0 : \rho = 0$ 에 대한 검정통계량

$$T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}} \sim t(n-2)$$

- 예 8.1 : 다음 자료는 어느 고등학교 학생 중에서 랜덤하게 추출된 20명의 언어영역과 외국어영역 점수이다. 언어영역과 외국어영역 성적이 이변량 정규분포를 따른다고 할 때, 이들 성적 사이의 상관관계가 있는지를 유의수준 5%에서 검정하여라

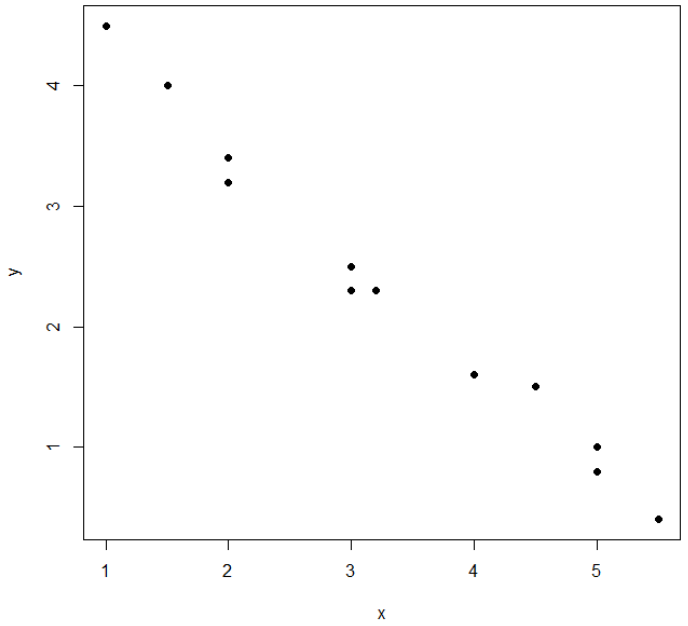
학생번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
언어영역	42	38	51	53	40	37	41	29	52	39	45	34	47	35	44	48	47	30	29	34
외국어영역	30	25	34	35	31	29	33	23	36	30	32	29	34	30	28	29	33	24	30	30

단순회귀분석

- 예 : 플라스틱 제품의 생산 공정에서 사출온도로부터 제품의 강도를 예측하는 경우
  - 설명 변수 (explanatory variable) : 다른 변수에 영향을 주는 변수 (사출온도)
  - 반응 변수 (response variable) : 영향을 받는 변수 (제품의 강도)

- 회귀분석의 목적은 설명 변수와 반응 변수의 관계를 구체적인 함수의 형태로 나타내고, 설명 변수의 값으로부터 반응변수의 값을 예측하는 것
- 회귀분석의 첫 단계는 산점도를 이용하여 두 변수의 관계를 파악하고 잠정적인 모형을 설정하는 것

- 예 : 사용년수(x)와 중고차가격(y)의 산점도
  - 산점도를 그려본 결과, 점들이 $y = \alpha + \beta x$ 주위에 분포되어 있음을 확인할 수 있음
  - 따라서 반응변수 y의 값은 주어진 x값에 대응되는 $\alpha + \beta x$ 값에 랜덤오차가 더해져서 나타나는 값으로 생각 할 수 있음



단순 선형 회귀모형

- 즉, 오차를 나타내는 확률변수 e 를 이용하면 반응변수 y 는 아래와 같은 확률변수 Y 의 관측값으로 이해할 수 있다

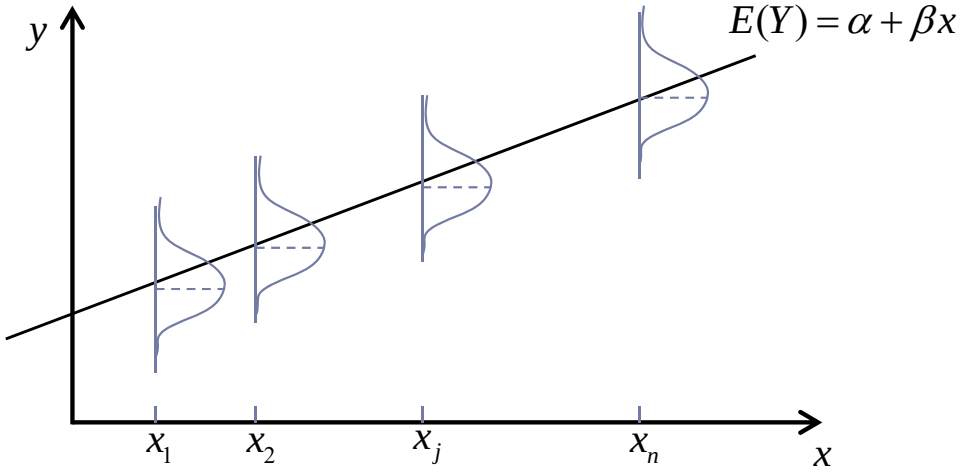
$$Y_i = \alpha + \beta x_i + e_i, \qquad e_i \sim ind.\ (0, \sigma^2)$$

- **단순선형 회귀모형**

$$Y_i = \alpha + \beta x_i + e_i \quad (i = 1, 2, ..., n)$$

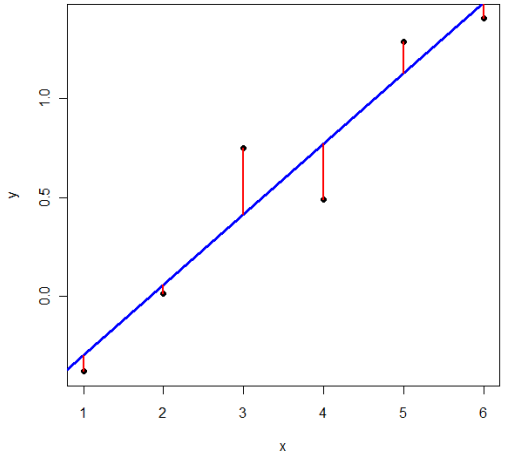
  - 선형성 : $E(e_i) = 0.$ 즉, $E(Y_i \mid x_i) = \alpha + \beta x_i$
  - 등분산성 : $Var(e_1) = ... = Var(e_n) = \sigma^2 > 0$
  - 독립성 : $e_1, ..., e_n$ 은 서로 독립

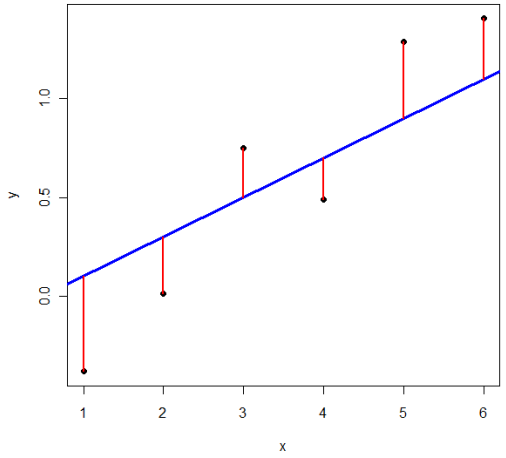
- 따라서 우리가 추정해야 할 직선은 $E(Y \mid x) = \alpha + \beta x$ 가 된다.



회귀계수의 추정

- Which is the best straight line?





- 추정된 직선 : $\hat{y}_i = \hat{E}(Y_i \mid x_i) = \hat{\alpha} + \hat{\beta} x_i$

- 잔차 (residual) : $\hat{e}_i = y_i - \hat{y}_i$
- 최소제곱법 (method of least squares)

  : 잔차제곱의 합을 최소로 만드는 직선을 추정

  : find $\hat{\alpha}, \hat{\beta}$ subject to minimize $Q = \sum_{i=1}^{n} \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{n} \{y_i - (\hat{\alpha} + \hat{\beta} x_i)\}^2$

최소제곱 회귀직선

- 최소제곱 추정량 (least squares estimator)

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}\ , \qquad \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{(xy)}}{S_{(xx)}}$$

- 최소제곱 회귀직선 (least squares regression line)

$$\hat{y} = \hat{\alpha} + \hat{\beta}x = \bar{y} + \hat{\beta}(x - \bar{x})$$

  - 여기서 $\hat{y} = \hat{E}(Y \mid x)$ 을 의미함.

- (참고) 최소제곱 추정량의 간편 계산식

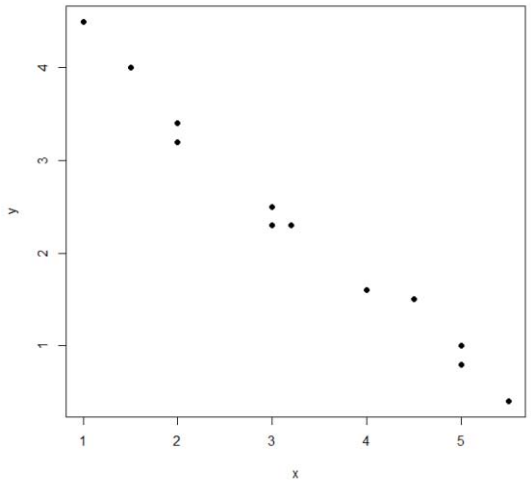
$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n = \sum x_i y_i - n\bar{x}\,\bar{y}$$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2/n = \sum x_i^2 - n\bar{x}^2$$

단순회귀직선의 추정

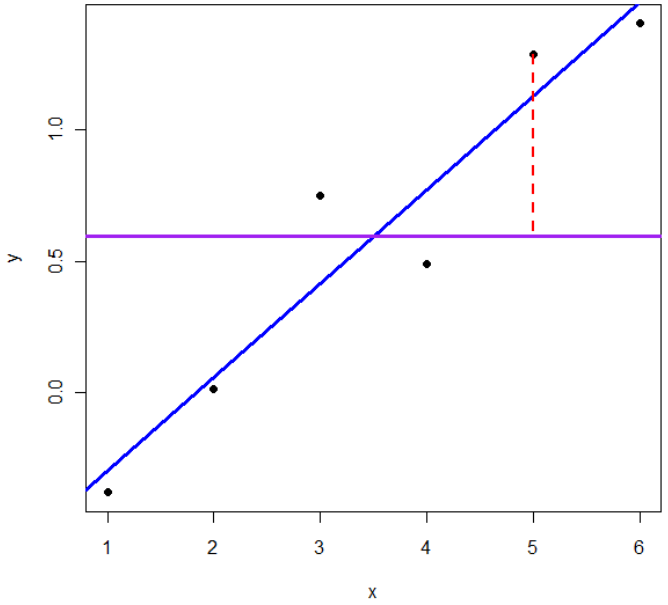
- 예제 8.2 : 사용년수에 따른 중고차 가격 자료를 이용하여 단순회귀직선을 추정하여라

사용년수	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4



회귀식의 설명력

- 총 편차의 분해 : 자료의 변동 $(y_i - \bar{y})$ 을 요인별로 분해



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

$$SST \quad = \quad SSR \quad + \quad SSE$$

SST = 총제곱합
: 전체 변동량

SSR = 회귀제곱합
: 회귀식으로 설명할 수 있는 부분

SSE = 잔차제곱합
: 회귀식으로 설명할 수 없는 부분

- 잔차제곱합 SSE는 오차분산 σ^2에 대한 정보를 제공하므로 오차분산 추정에 이용됨

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

결정계수

- **결정계수(coefficient of determination)**

  : 회귀식의 설명력을 나타내는 지표

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 결정계수의 성질
  - 결정계수는 0에서 1 사이의 값을 갖는다
  - 1에 가까울수록 회귀식의 설명력이 높다는 것을 의미
  - 결정계수가 1에 가까울수록 데이터들이 회귀직선 주위로 밀집되어 나타남

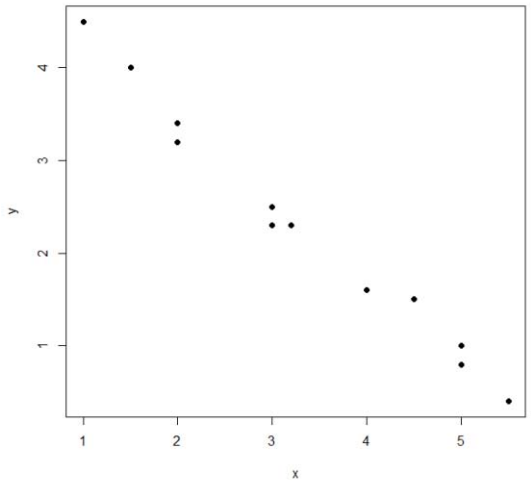
$$r^2 = \frac{SSR}{SST} = \frac{\{S_{(xy)}\}^2}{S_{(xx)}S_{(yy)}} = \left\{ \frac{S_{(xy)}}{\sqrt{S_{(xx)}}\sqrt{S_{(yy)}}} \right\}^2$$

  - 결정계수가 높을수록 좋지만, 기준이 되는 값은 없음. 일반적으로 0.6이상이면 설명력이 좋은 편이라고 생각할 수 있음
  - 결정계수는 설명변수가 증가하면 함께 증가하는 경향이 있음. 따라서 결정계수가 절대적인 모형의 평가 기준이 될 수는 없음

결정계수

- 예제 8.2 : 사용년수에 따른 중고차 가격 자료를 이용하여 단순회귀모형을 적용할 때, 결정계수를 구하고 이를 해석하여라

사용년수	1.0	1.5	2.0	2.0	3.0	3.0	3.2	4.0	4.5	5.0	5.0	5.5
중고차가격	4.5	4.0	3.2	3.4	2.5	2.3	2.3	1.6	1.5	1.0	0.8	0.4



단순회귀분석에서의 추론

- 단순선형회귀모형의 기본 가정
  : 선형성, 등분산성, 독립성, 정규성 가정이 만족하게 되면 추론이 가능

$$\begin{cases} Y_i = \alpha + \beta x_i + e_i \\ e_i \sim ind.\ N(0, \sigma^2) \end{cases}$$

  : 잔차분석을 통해 기본 가정의 타당성 검토

- 회귀직선의 유의성 검정
  : 모회귀계수 β에 관한 구간추정과 가설 검정

- 모회귀계수 α에 관한 추론

- 평균반응 $E(Y \mid x) = \alpha + \beta x$에 관한 추론

회귀직선의 유의성 검정

- 최소제곱법을 통해 찾아낸 회귀식이 과연 의미있는 것인지를 확인하는 작업
  - 회귀직선의 설명력을 이용하여 검정하는 방법 (F-test)
  - 회귀계수의 추정량의 분포를 이용하여 검정하는 방법 (t-test)

- F-test
  - 가설 : $H_0 : \beta = 0 \quad vs \quad H_1 : \beta \neq 0$
  - 회귀직선이 유의하다면, SSR의 비중이 높아질 것임
  - 즉, SSR/SSE의 값이 커질수록 회귀직선이 유의하다는 증거가 강해지게 됨
  - 따라서 아래의 분포를 이용하여 가설검정을 할 수 있음

$$F = \frac{SSR/1}{SSE/n-2} \sim F(1,\ n-2)$$

  - 이러한 유의성 검정 결과는 아래와 같은 분산분석표의 형태로 요약할수 있음

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	SSR	1	MSR	f= MSR/MSE	$P(F \geq f)$
잔차	SSE	n-2	MSE		
전체	SST	n-1			

회귀직선의 유의성 검정

- T-test
  - 모회귀계수의 추정량 $\hat{\beta}$ 의 표본 분포

$$\hat{\beta} \sim N\left(\beta,\ \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

  - 모회귀계수 β에 관한 $100(1-\alpha)\%$ 신뢰구간

$$\left[\hat{\beta} - t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{(xx)}}},\quad \hat{\beta} + t_{\alpha/2}(n-2)\frac{\hat{\sigma}}{\sqrt{S_{(xx)}}}\right]$$

  - 귀무가설 $H_0 : \beta = b$ 에 관한 검정통계량

$$\frac{\hat{\beta} - b}{\hat{\sigma} / \sqrt{S_{(xx)}}} \sim t(n-2)$$

where $\hat{\sigma}^2 = MSE$

회귀직선의 유의성 검정

- 예제 8.5 : 예 8.2의 사용년수에 따른 중고차 가격 자료를 이용하여 단순회귀모형을 적용할 때, 회귀직선의 유의성을 F-test와 t-test를 이용하여 유의수준 1%에서 각각 검정하시오. 두 검정 사이에는 어떠한 관련성이 있는가?

평균반응에 관한 추론

- 평균반응 $E(Y \mid x) = \alpha + \beta x$ 의 추정량의 표본 분포

$$\hat{\alpha} + \hat{\beta} x \sim N\left(\alpha + \beta x,\ \ \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{(xx)}} \right) \right)$$

- 평균반응 $\alpha + \beta x$ 에 관한 $100(1-\alpha)\%$ 신뢰구간

$$\left[(\hat{\alpha} + \hat{\beta} x) - t_{\alpha/2}(n-2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{(xx)}}},\quad (\hat{\alpha} + \hat{\beta} x) + t_{\alpha/2}(n-2)\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{(xx)}}} \right]$$

- 귀무가설 $H_0 : \alpha + \beta x = \mu_0$ 에 관한 검정통계량

$$\frac{(\hat{\alpha} + \hat{\beta} x) - \mu_0}{\hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{(xx)}}}} \ \sim\ t(n-2)$$

모회귀직선의 절편에 관한 추론

- 모회귀계수 α 의 추정량의 표본 분포

$$\hat{\alpha} \sim N\left(\alpha,\ \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{(xx)}}\right)\right)$$

- 모회귀계수 α 에 관한 $100(1-\alpha)\%$ 신뢰구간

$$\left[\hat{\alpha}-t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{(xx)}}},\quad \hat{\alpha}+t_{\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{(xx)}}}\right]$$

- 귀무가설 $H_0:\alpha=a$ 에 관한 검정통계량

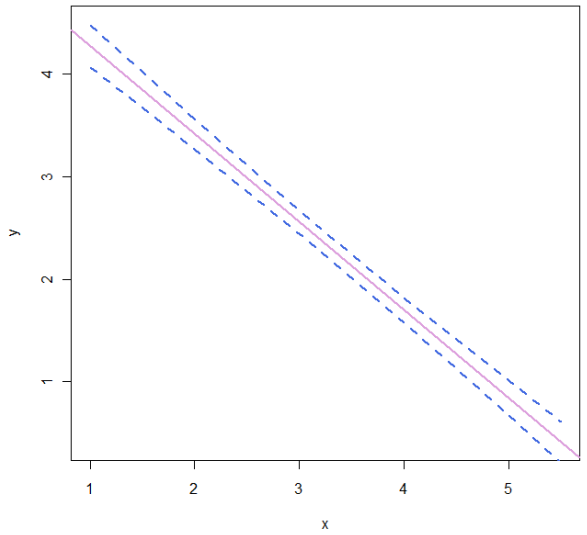
$$\frac{\hat{\alpha}-a}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{S_{(xx)}}}} \sim t(n-2)$$

회귀직선에 관한 추론 : 예

- 예 8.7 : 예 8.2에서 단순선형회귀 모형을 적용할 때,

(1) 사용년수가 2.5년인 중고차의 평균 가격이 2,800,000원보다 높은지 유의수준 1%에서 검정하시오

(2) 사용년수가 x=1,3,5 일 때 중고차의 평균 가격의 95% 신뢰구간을 각각 구하시오



잔차 분석

- 단순회귀모형의 가정사항

$$Y_i = \alpha + \beta x_i + e_i, \qquad i = 1, 2, ..., n.$$

  - (a) 선형성 : $E(e_i) = 0, \quad E(Y_i \mid x_i) = \alpha + \beta x_i$
  - (b) 등분산성 : $Var(e_i) = \sigma^2 > 0, \quad \forall i = 1, 2, ..., n.$
  - (c) 독립성 : $e_1, e_2, ..., e_n$ 은 서로 독립
  - (d) 정규성 : e_i 는 정규분포를 따름

- 오차의 관측값인 잔차를 이용하여 가정을 검토할 수 있음

- **잔차분석 (analysis of residual)**
  : 스튜던트화 잔차를 이용하여 단순선형회귀모형의 가정의 타당성 여부 검토

$$\hat{e}_{st,i} = \frac{\hat{e}_i}{\widehat{sd}(\hat{e}_i)} = \frac{\{y_i - \hat{y}_i\}}{\widehat{sd}(\hat{e}_i)}$$

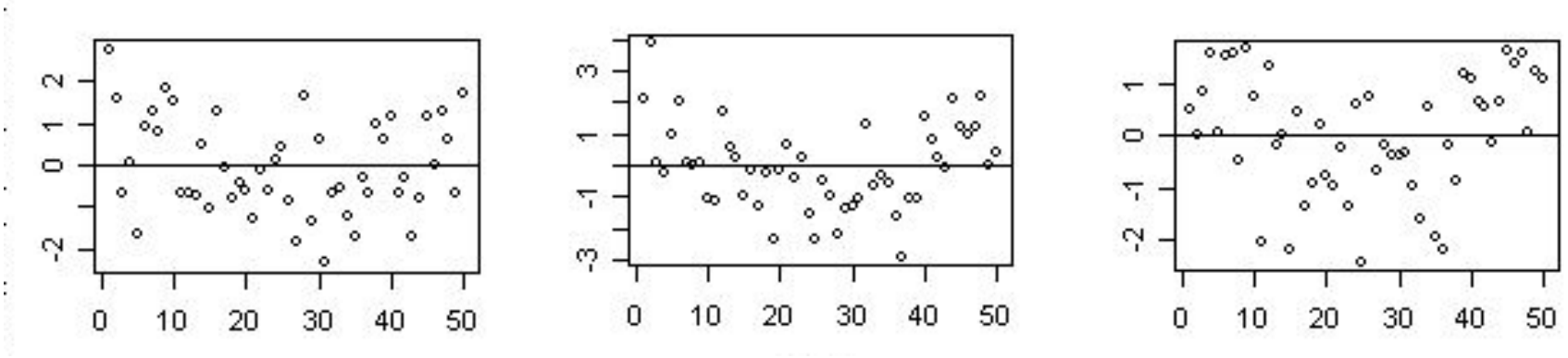
잔차도

- 잔차도(residual plot)
  : 설명변수와 잔차를 산점도로 나타낸 것
  : 잔차도에서 스튜던트화 잔차들이
  (1) 대략 0에 관하여 대칭적으로 나타나고
  (2) 설명변수 값에 따른 잔차의 산포가 크게 다르지 않고
  (3) 점들이 특정한 형식을 가지고 나타남이 없으며
  (4) 모든 점이 (-2,2) 범위 내에 나타나게 되면
  오차에 대한 가정이 만족함을 알 수 있다.

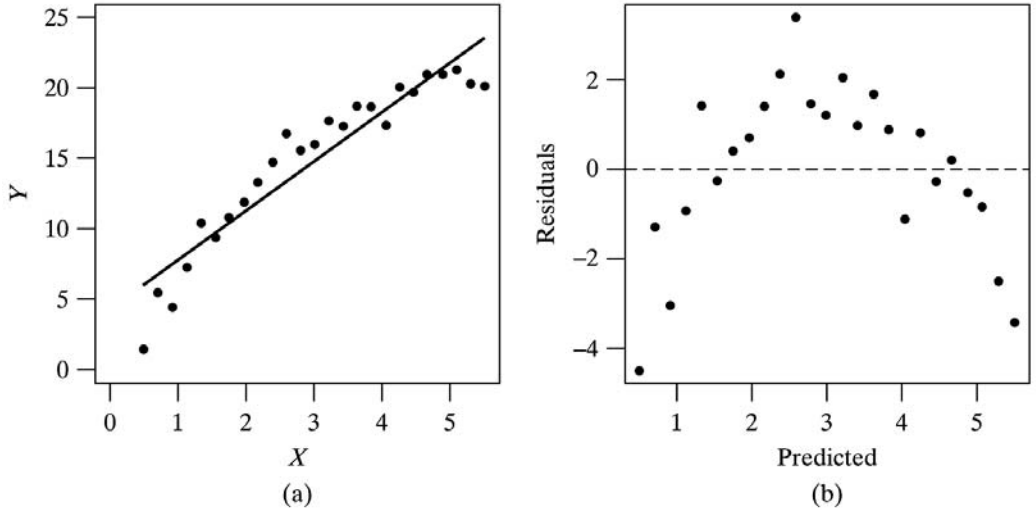
- 잔차는 회귀식으로 설명되지 않는 부분이므로 정보를 포함하고 있어서는 안됨
- 만약 잔차도에서 어떠한 패턴(곡선 모양 또는 증가, 감소의 경향 등)이 나타난다면, 설명 가능한 정보가 버려졌다는 의미이므로 회귀식이 잘못 적합 되었음을 의미함

잔차도

- 다양한 잔차도의 형태



- 단순선형회귀모형 가정이 어긋나는 경우의 잔차도



중회귀분석

- 중회귀분석 (multiple regression)
  : 반응변수의 변화를 설명하기 위하여 두 개 이상의 설명변수를 고려하는 회귀분석

- 중회귀모형

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_i, \qquad i = 1, 2, \ldots, n.$$

  where $e_i \sim N(0, \sigma^2)$ 이고 서로 독립.

- 중회귀선형모형의 계수 추정 역시, 잔차제곱합을 최소로 하는 계수를 찾는 방법인 최소제곱법을 사용한다

중회귀분석

- 중회귀분석의 제곱합 분해

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

$$SST \quad = \quad SSE \quad + \quad SSR$$

$$(n-1) \quad = \quad (n-k-1) + \quad (k)$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki}$

- 오차분산의 추정량 : $\hat{\sigma}^2 = \frac{SSE}{n-k-1} = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

- 결정계수 : $r^2 = \frac{SSR}{SST}$

- (참고) 수정된 결정계수(adjusted R-square)

  : 모형 추정에 사용 된 설명변수의 개수에 따라 적절한 패널티를 부과

  : 수정된 결정계수를 이용하면, 유의하지 않은 설명변수가 많이 추가 된 다중회귀모형의 경우에는 오히려 설명력이 감소

$$r_a^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \frac{SSE}{SST}$$

중회귀분석의 유의성 검정

- **F-검정 : overall significance test, 전체적 검정**
  - 추정된 회귀식 전체의 유의성을 검정하는 방법
  - 각 계수별 유의성이 아닌, 모형 전체의 전반적인 유의성을 검정하는 방법
  - 가설 설정 : $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

    $H_1 : \text{Not } H_0$

- **t-검정 : individual significance test, 개별적 검정**
  - 추정된 회귀식의 계수들의 유의성을 검정하는 방법
  - 각 계수별 유의성을 개별적으로 검정하는 방법
  - 가설설정 : $H_0 : \beta_i = 0$

    $H_1 : \text{Not } H_0$

- 단순회귀분석에서는 두 검정이 같은 의미를 지님

회귀식의 유의성 검정 : F-test

- 모회귀함수의 유의성 검정

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : Not \quad H_0.$$

- 검정통계량

$$F = \frac{MSR}{MSE} \sim F(k, n-k-1)$$

- 중회귀분석의 분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	SSR	k	MSR	f= MSR/MSE	$P(F \geq f)$
잔차	SSE	n-k-1	MSE		
전체	SST	n-1			

중회귀분석 : 예제

- 어떤 플라스틱 제품 생산공정에서 제품의 수율은 그 제품을 만들때 소요되는 공정시간과 원료의 촉매량에 영향을 받는다고 한다. 다음의 자료를 이용하여 중회귀분석을 실시한 결과가 아래와 같다. 결과를 이용하여 모회귀함수의 유의성검정을 유의수준 5%에서 하고, 결정계수의 값을 구하시오.

번호	1	2	3	4	5	6	7	8	9	10
X1 (분)	8	8	10	10	12	12	14	14	16	16
X2 (g)	3.1	3.3	3.5	3.0	3.2	3.4	3.0	3.6	3.2	3.6
Y (%)	75	80	84	77	79	85	86	90	87	89

분산분석표

요인	제곱합	자유도	평균제곱	F 값	유의확률
회귀	204.7021	2	102.3511	20.530	0.0012
잔차	34.8979	7	4.9854		
계	239.6000	9			

모수 추정

변수명	추정값	표준오차	t-통계량	유의확률
절편	36.11144	10.75101	3.359	0.0121
x1	1.20158	0.26117	4.601	0.0025
x2	9.92999	3.41095	2.911	0.0226