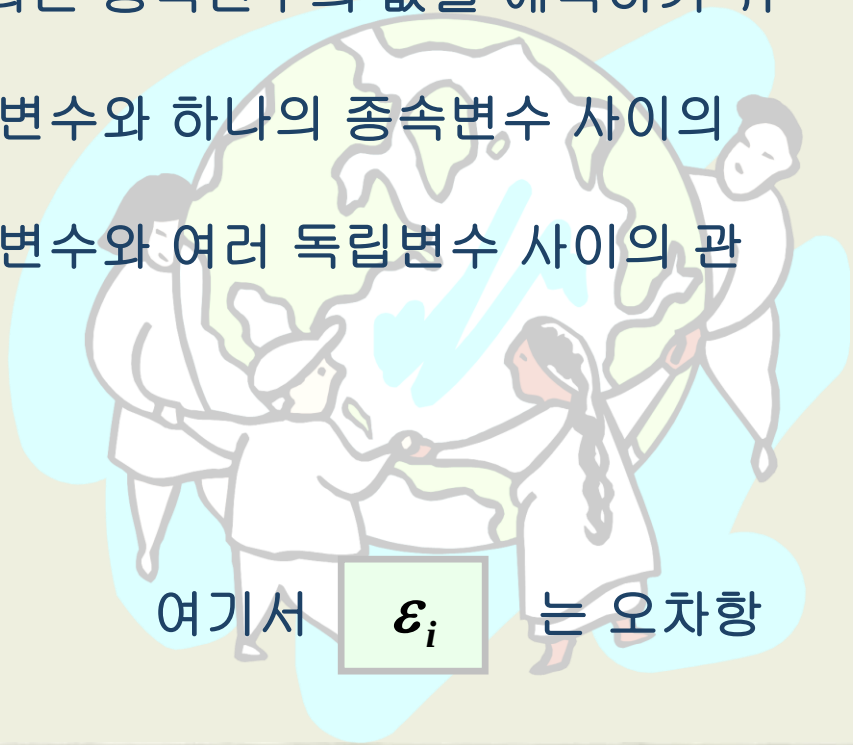
제 13/14장

회귀분석/다중회귀모형

고려대학교 경영대학 박 광태

회귀분석

- 회귀분석
  - 독립변수가 종속변수에 미치는 영향력의 크기를 측정하여 독립변수의 일정한 값에 대응되는 종속변수의 값을 예측하기 위한 방법
  - 단순회귀분석 : 하나의 독립변수와 하나의 종속변수 사이의 관계를 분석.
  - 다중회귀분석 : 하나의 종속변수와 여러 독립변수 사이의 관계를 분석.



- 모집단 회귀식
  - $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

여기서 ε_i 는 오차항

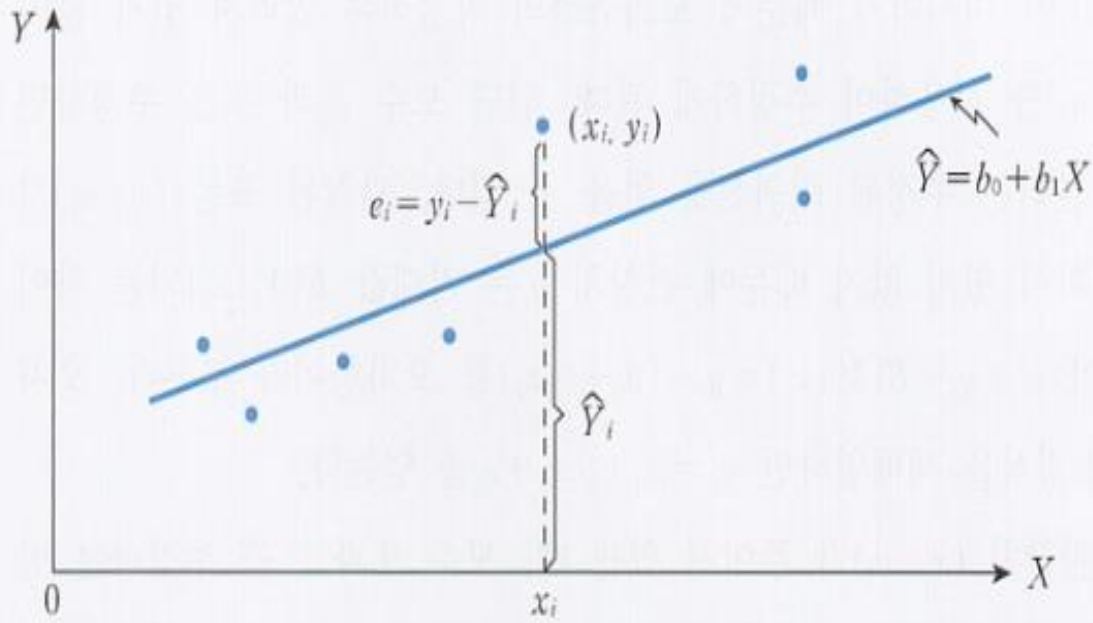
회귀분석

- 표본회귀식
  - $\hat{Y}_i = b_0 + b_1 x_i$ 여기서 잔차 $e_i = Y_i - \hat{Y}_i$
- 표본회귀식을 구하는 방법(최소제곱법)
  - 잔차들의 제곱합(SSE)이 최소가 되게 회귀식을 구하는 방법
- 회귀모형의 기본가정

$$1)\ E(e_i) = 0$$
$$2)\ Var(e_i) = \sigma_e^2$$
$$3)\ E(e_i e_j) = 0$$
$$4)\ E(X_i e_i) = 0$$
$$5)\ e_i \sim N(0, \sigma_e^2)$$

회귀분석

그림 13-3 y_i, $\hat{Y}_i$와 잔차 e_i 사이의 관계를 보여 주는 도표



회귀분석

그림 13-4 등분산오차를 갖는 모회귀선

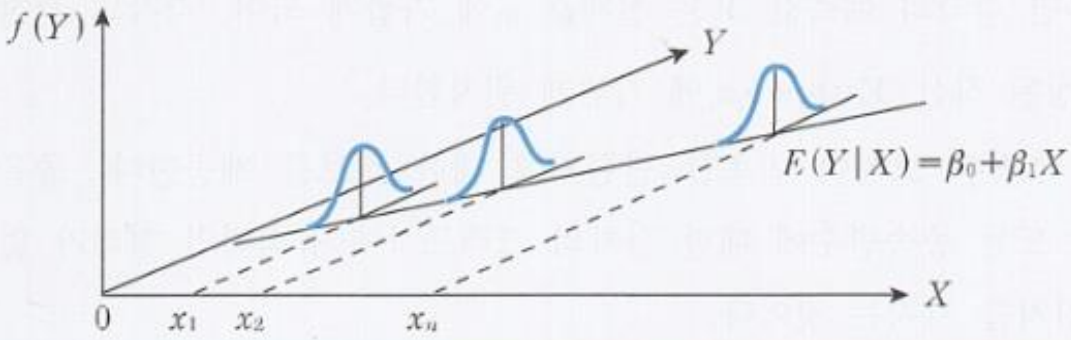
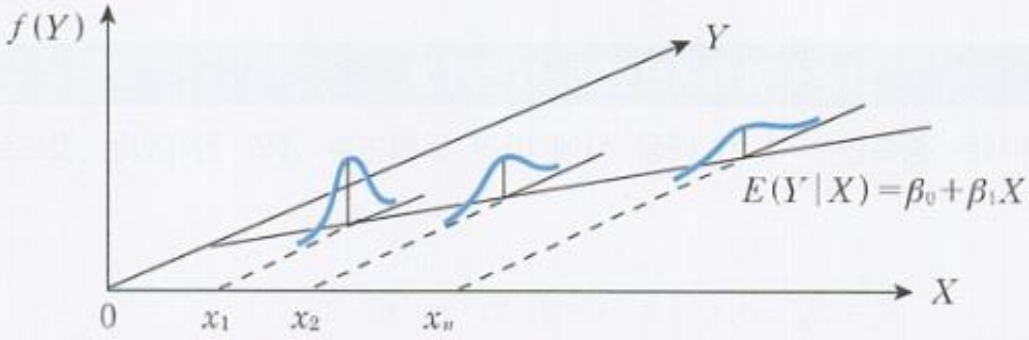


그림 13-5 이분산오차를 갖는 모회귀선



예제 13-2

천마기업은 월 매출액(단위: 백만 원)이 월 광고비지출액(단위: 백만 원)에 대략 선형으로 관계된다고 생각한다. 과거 10개월의 자료가 표 13-1과 그림 13-6에 주어져 있다. 표본회귀선을 계산하고 $X=13$ 일 때, Y의 예측값을 구하시오.

표 13-1 월간 광고비지출액과 매출액 자료 (단위: 백만 원)

월 i	월간 광고비 지출액 x_i	월간 매출액 y_i	x_iy_i	x_i^2	y_i^2
1	3	40	120	9	1,600
2	4	50	200	16	2,500
3	5	45	225	25	2,025
4	4	45	180	16	2,025
5	5	50	250	25	2,500
6	6	55	330	36	3,025
7	7	70	490	49	4,900
8	8	85	680	64	7,225
9	12	100	1,200	144	10,000
10	13	115	1,495	169	13,225
합 계	67	655	5,170	553	49,025

표 13-6 표 13-1의 자료에 대한 산포도

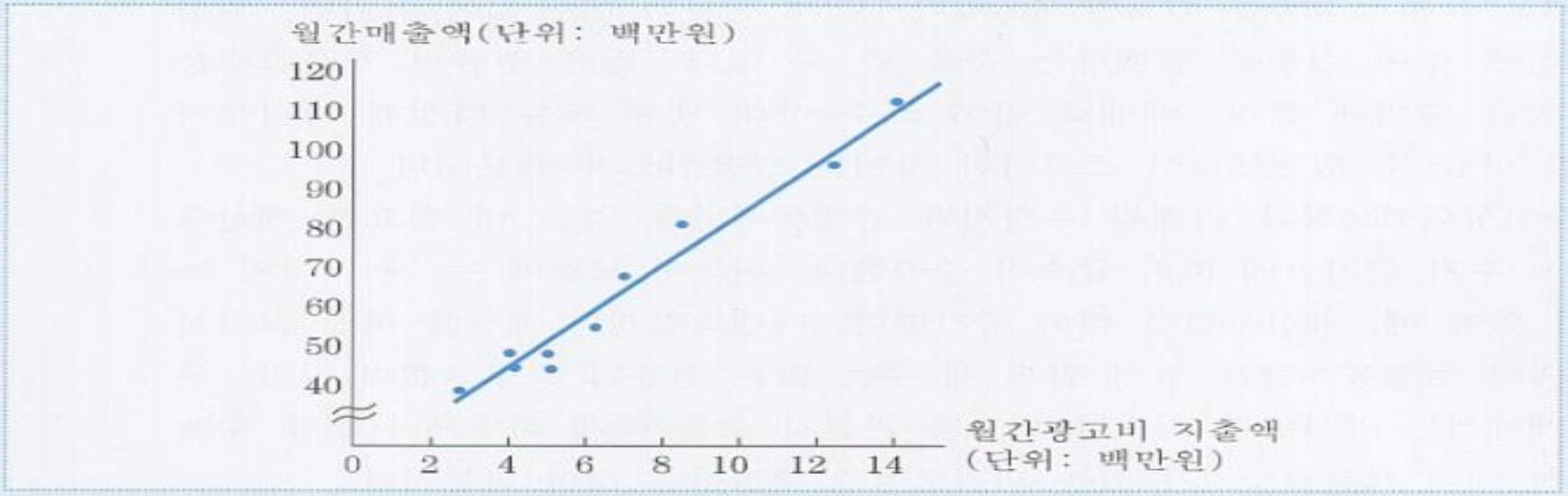
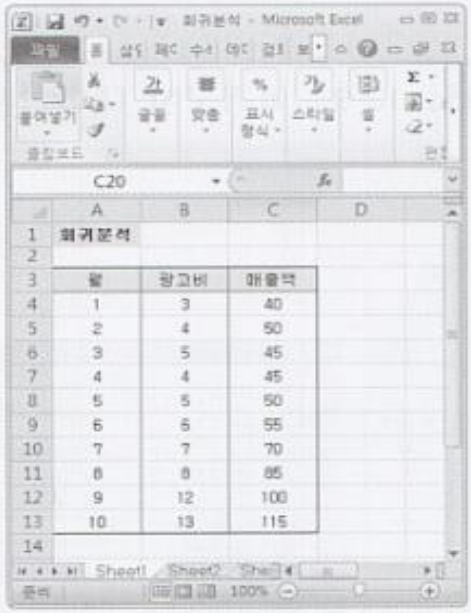


그림 13-7과 같이 예제 2의 자료를 입력한다.

그림 13-7 자료입력



회귀분석을 위해 데이터 메뉴에서 데이터 분석을 선택하면 그림 13-8의 화면이 나타난다. 여기서 『회귀분석』을 선택하고 확인을 누르면 그림 13-9의 회귀분석 대화상자가 나타나게 된다.

그림 13-8 회귀분석 선택

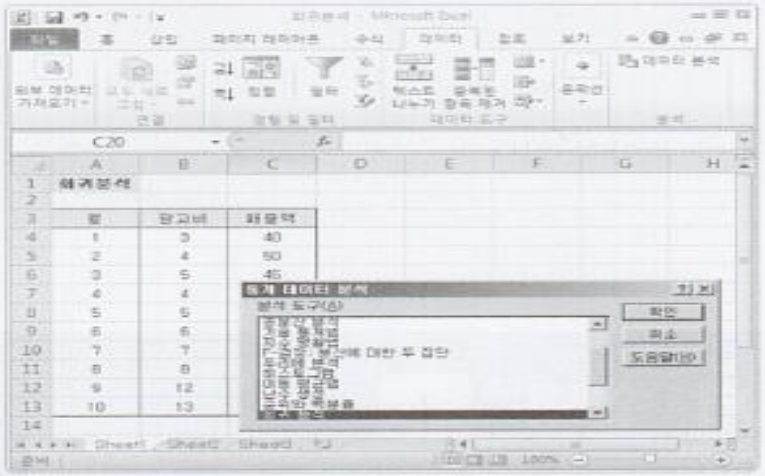


그림 13-9에서 Y축은 매출액자료이고 X축은 광고비자료이므로 해당되는 셀의 범위를 대화상자에 입력한다. 이름표에 표시하고 출력범위는 E2셀로 지정한 후 확인을 누르면 그림 13-10과 같이 결과가 주어진다.

그림 13-9 회귀분석 대화상자

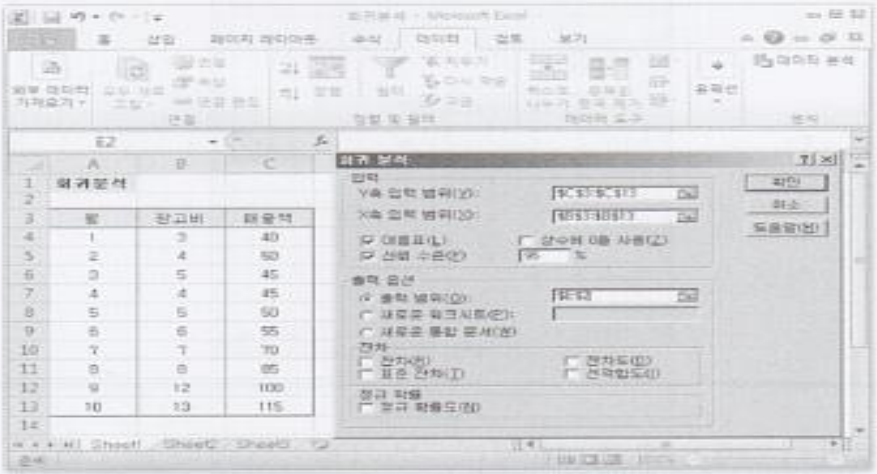
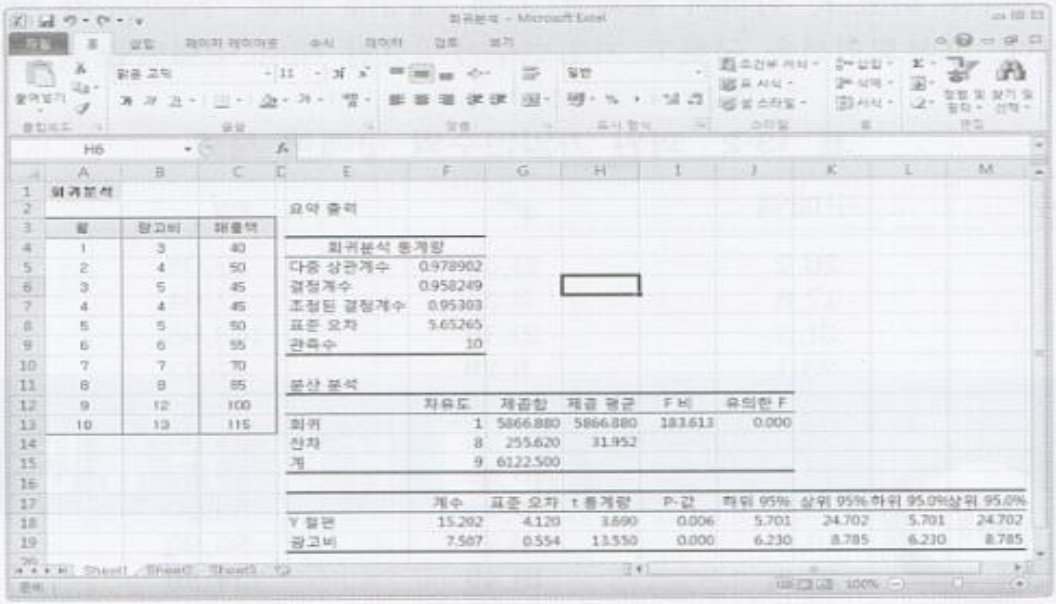


그림 13-10 회귀분석의 결과



회귀분석

월	광고비	매출액
1	3	40
2	4	50
3	5	45
4	4	45
5	5	50
6	6	55
7	7	70
8	8	85
9	12	100
10	13	115

요약 출력

회귀분석 통계량	
다중 상관계수	0.978902
결정계수	0.958249
조정된 결정계수	0.95303
표준 오차	5.65265
관측수	10

분산 분석

	자유도	제곱합	제곱 평균	F 비	유의한 F
회귀	1	5866.880	5866.880	183.613	0.000
잔차	8	255.620	31.952		
계	9	6122.500			

	계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
Y 절편	15.202	4.120	3.690	0.006	5.701	24.702	5.701	24.702
광고비	7.507	0.554	13.550	0.000	6.230	8.785	6.230	8.785

그림 13–10의 결과를 살펴보면, 결정계수가 0.958로 주어지는데 이는 종속변수의 변동 중 95.8%가 독립변수(여기서는 광고비지출액)에 의해 설명됨을 의미한다. 그리고 분산분석의 유의한 F하의 값, 즉 p값이 0.0000으로 유의수준 1%보다 낮은 유의수준에서도 추정된 회귀식이 유의함을 알 수 있다. 끝으로 추정된 회귀식은 Y절편이 15.2017이고 독립변수인 광고비지출액의 계수가 7.5072인 $\hat{Y}=15.2017+7.5072X$로 표시됨을 알 수 있다. 여기서 광고비지출액의 계수에 대한 p값이 0.01보다 작기 때문에 1% 유의수준에서 매출액에 대한 광고비지출액의 회귀계수가 통계적으로 유의함을 알 수 있다. 즉, 독립변수가 종속변수를 설명하는 변수로 의미가 있음을 알 수 있다. 또한 회귀식에서 기울기가 양(+)의 값을 갖는데 이는 두 변수가 정(+)의 상관관계를 가짐을 의미한다.

예제 13-4

예제 2에 대해 회귀의 추정된 표준오차 S_e를 계산하시오.

제곱합 등식

- 제곱합 등식

총변동=설명변동+비설명변동

$$SST = SSR + SSE$$

$$SST = \sum (Y_i - \bar{Y})^2$$

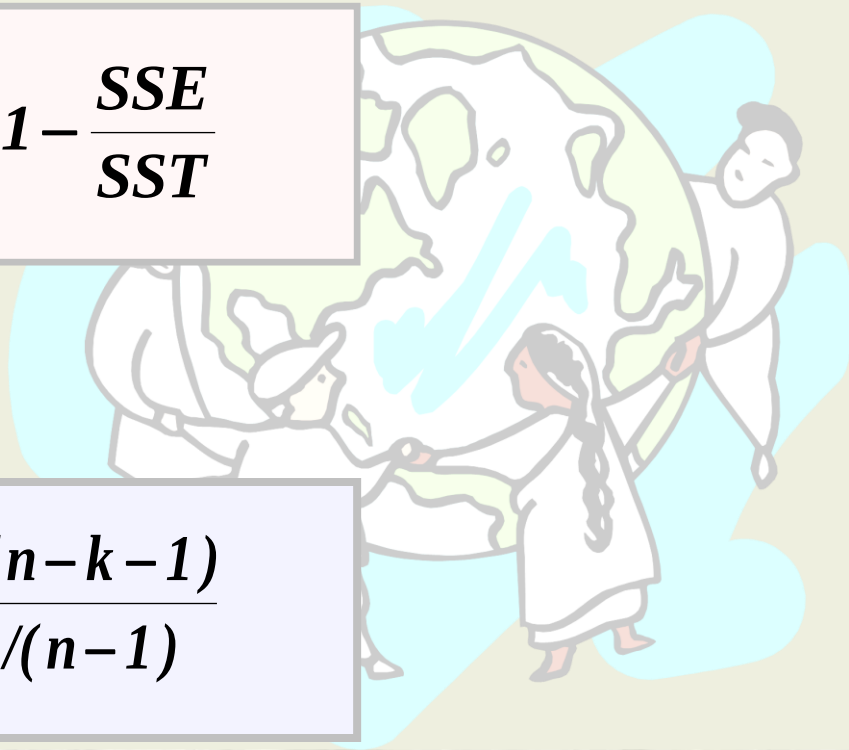
$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

결정계수

- 결정계수
  - 종속변수의 변동을 독립변수가 얼마만큼 설명해주는 가를 나타냄

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



- 조정결정계수

$$R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

예제 13-5

예제 2에서 추정된 매출액($\hat{Y}_i$)과 광고비지출액(x_i)을 관계시키는 다음의 표본회귀방정식을 얻었다.

$$\hat{Y}_i = 15.2017 + 7.5072x_i$$

또한 예제 4에서 모든 잔차와 SSE를 얻었다. 이 모형에 대한 R^2을 계산하시오.

ANOVA 표

- ANOVA 표의 작성

변동요인	제곱합	자유도	평균제곱합	F -ratio
회귀	SSR	k	$MSR=SSR/k$	MSR/MSE
잔차	SSE	$(n-k-1)$	$MSE=SSE/(n-k-1)$	
합	SST	$(n-1)$		

- 회귀모형의 유의성 검정에 이용됨

- 표준오차

$$S_e = MSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}}$$

회귀 모형 가설검정 & 신뢰구간 추정

- 단일모수 β에 대한 가설검정

$$t_{\alpha/2,\,n-k-1} = \frac{b_i - 0}{S_b}$$

- 단일모수 β의 신뢰구간 추정

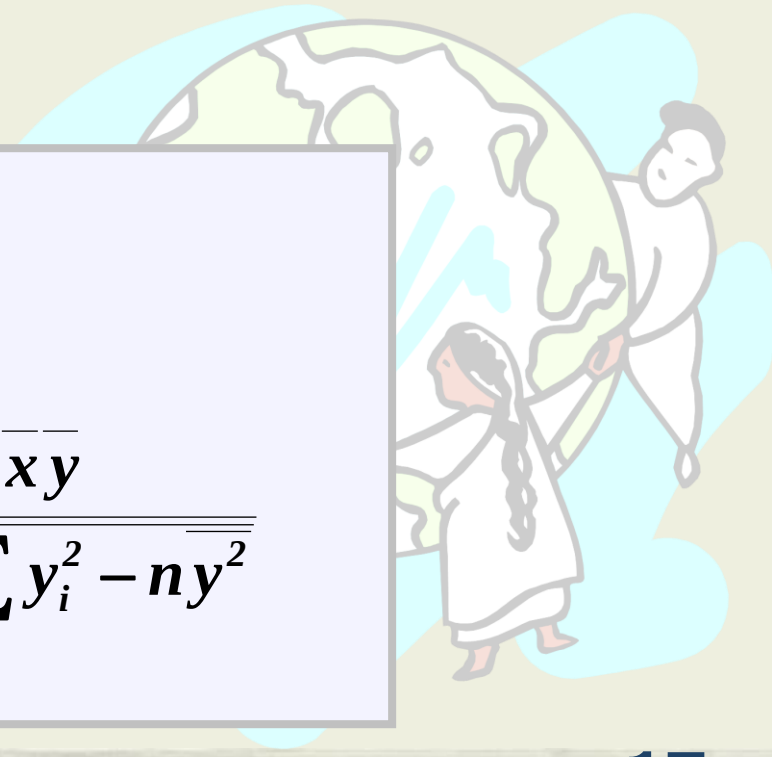
$$b_i - t_{\alpha/2,\,n-k-1} S_b \le \beta_i \le b_i + t_{\alpha/2,\,n-k-1} S_b$$

상관분석

- 상관분석
  - 변수들과의 관계의 강도, 즉 변수들이 얼마나 밀접하게 관련되어 있는가를 분석하는 것

- 표본상관계수

$$R = \frac{S_{XY}}{S_X S_Y} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x^2}}\sqrt{\sum y_i^2 - n\bar{y^2}}}$$



예제 13-10

예제 2를 참고하여 광고비지출액과 매출액 사이의 상관계수를 계산하시오.

- 다중회귀모형

: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$

- 모회귀방정식

: $E(Y_i | x_{i1}, x_{i2}, \cdots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$

- 표본회귀방정식

: $\hat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}$

- 최소제곱추정치

: 모수 $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$의 최소제곱추정치는 잔차제곱합 SSE를 최소로 하는 값 $b_0, b_1, b_2, \cdots, b_k$이다.

여기서

$SSE = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik})^2$이다.

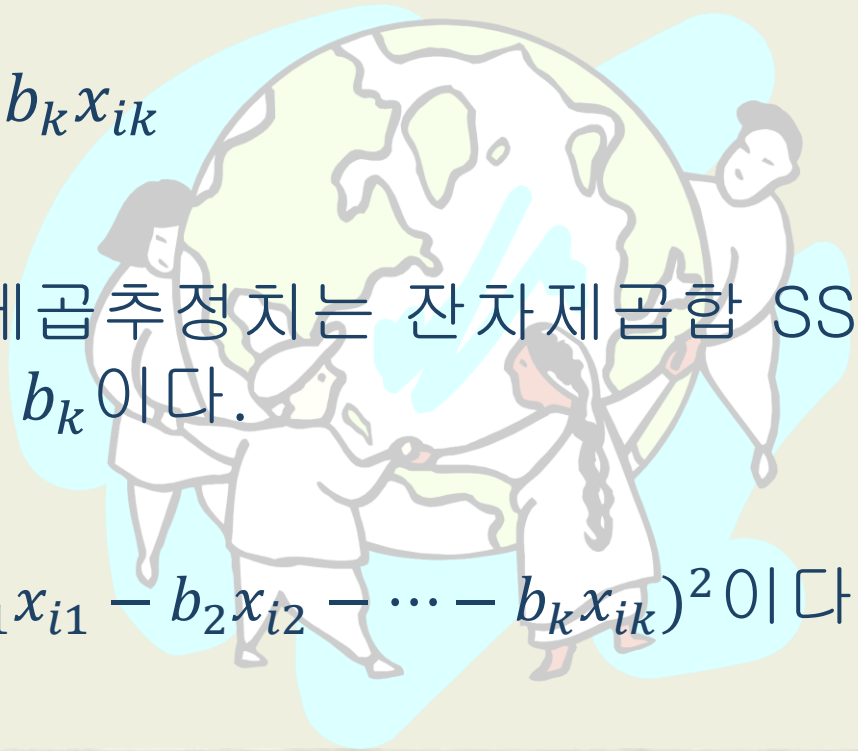
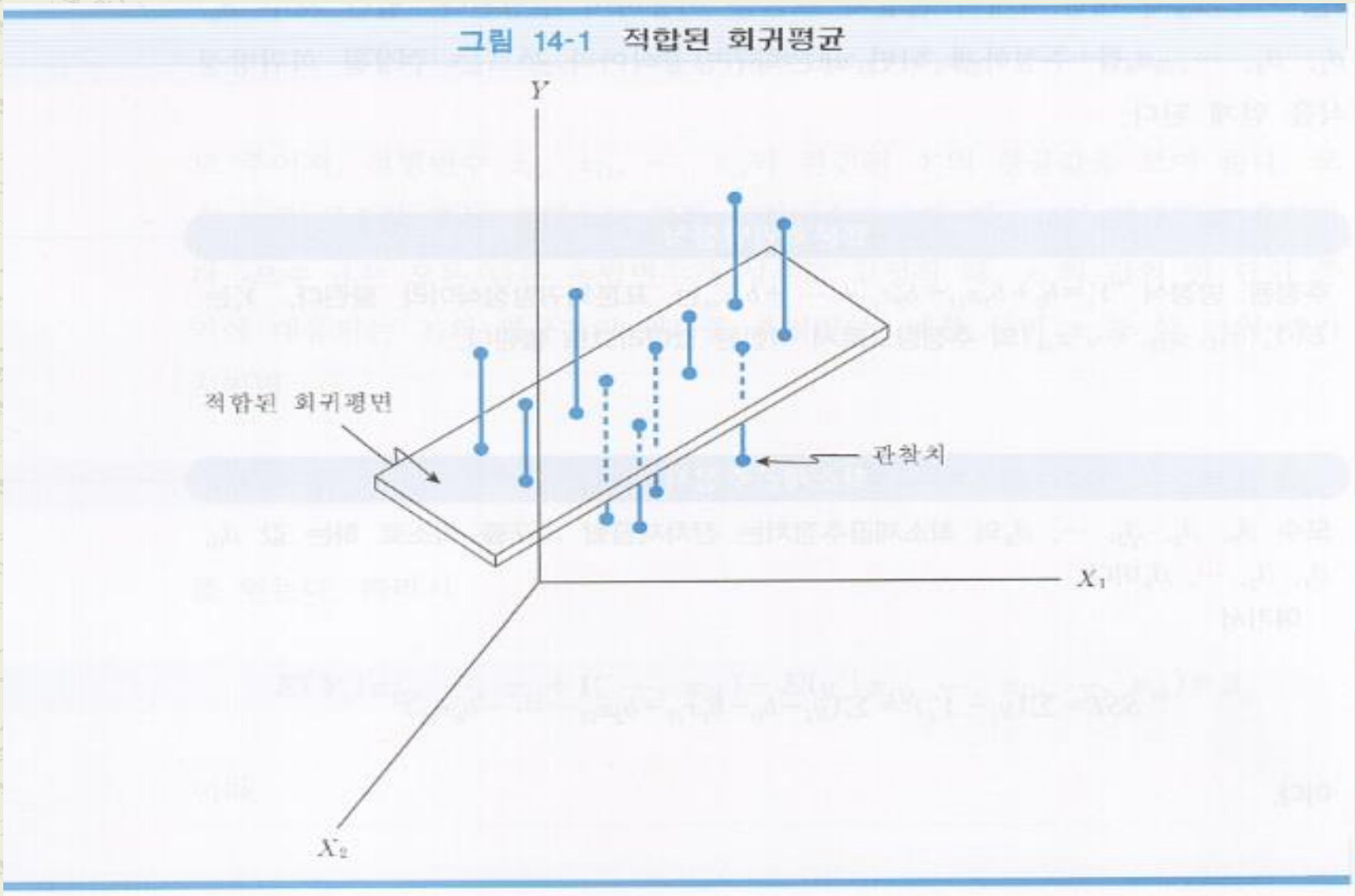


그림 14-1



두 개의 설명변수를 포함한 표본회귀식은 $\hat{Y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2}$이다. 최소제곱법에 따르면 아래의 잔차제곱합 SSE를 최소로 하는 추정치 b_0, b_1과 b_2를 선택하게 된다.

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

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

추정량 b_0, b_1과 b_2를 찾기 위해 SSE를 b_0, b_1과 b_2에 대해 편미분하여 얻은 다음의 정규방정식(normal equation)을 동시에 만족시키는 b_0, b_1과 b_2를 구한다.

$$\Sigma y_i = nb_0 + b_1 \Sigma x_{i1} + b_2 \Sigma x_{i2}$$

$$\Sigma x_{i1} y_i = b_0 \Sigma x_{i1} + b_1 \Sigma x_{i1}^2 + b_2 \Sigma x_{i1} x_{i2}$$

$$\Sigma x_{i2} y_i = b_0 \Sigma x_{i2} + b_1 \Sigma x_{i1} x_{i2} + b_2 \Sigma x_{i2}^2$$

즉,

$$b_1 = \frac{(\Sigma x_{i2}^2)(\Sigma x_{i1} Y_i) - (\Sigma x_{i1} x_{i2})(\Sigma x_{i2} Y_i)}{(\Sigma x_{i1}^2)(\Sigma x_{i2}^2) - (\Sigma x_{i1} x_{i2})^2}$$

$$b_2 = \frac{(\Sigma x_{i1}^2)(\Sigma x_{i2} Y_i) - (\Sigma x_{i1} x_{i2})(\Sigma x_{i1} Y_i)}{(\Sigma x_{i1}^2)(\Sigma x_{i2}^2) - (\Sigma x_{i1} x_{i2})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{x_1} - b_2 \overline{x_2}$$

이다.

예제 14-1

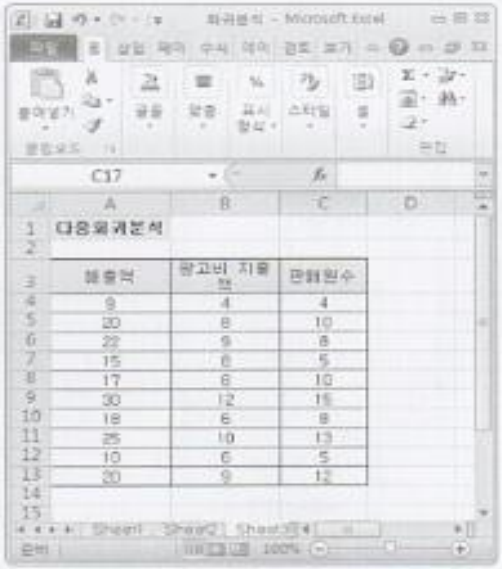
대리점의 매출액(단위: 천만 원)은 광고비지출액(단위: 백만 원)과 판매원 수(단위: 명)에 의해 결정된다고 볼 수 있다. $n=10$개의 대리점 표본에 대한 정보가 표 14-1에 주어져 있다. 즉, 표 14-1의 자료는 i번째 대리점의 매출액 y_i, i번째 대리점의 광고비지출액 x_{i1} 그리고 i번째 대리점의 판매원 수 x_{i2}를 보여 준다. 회귀평면을 추정하고, 광고비지출액 10과 판매원 수 16명을 갖는 대리점에 대한 매출액을 예측하시오.

표 14-1 예제 1에 대한 자료

매출액 (단위: 천만 원) Y	광고비지출액 (단위: 백만 원) X_1	판매원 수 (단위: 명) X_2	X_1^2	X_2^2	X_1Y	X_2Y	X_1X_2	Y^2
9	4	4	16	16	36	36	16	81
20	8	10	64	100	160	200	80	400
22	9	8	81	64	198	176	72	484
15	8	5	64	25	120	75	40	225
17	8	10	64	100	136	170	80	289
30	12	15	144	225	360	450	180	900
18	6	8	36	64	108	144	48	324
25	10	13	100	169	250	325	130	625
10	6	5	36	25	60	50	30	100
20	9	12	81	144	180	240	108	400
합계 186	80	90	686	932	1,608	1,866	784	3,828

그림 14-2와 같이 예제 1의 자료를 입력한다.

그림 14-2 자료입력



	A	B	C	D
1	다중회귀분석			
2				
3	매출액	광고비 지출액	판매원수	
4	9	4	4	
5	20	8	10	
6	22	9	8	
7	15	6	5	
8	17	8	10	
9	30	12	15	
10	18	6	8	
11	25	10	13	
12	10	6	5	
13	20	9	12	
14				
15				

이제 데이터 메뉴에서 데이터 분석을 선택하면 그림 14-3의 화면이 나타난다. 여기서 『회귀분석』을 선택하고 확인을 누르면 그림 14-4의 회귀분석 대화상자가 나타나게 된다.

그림 14-3 회귀분석 선택

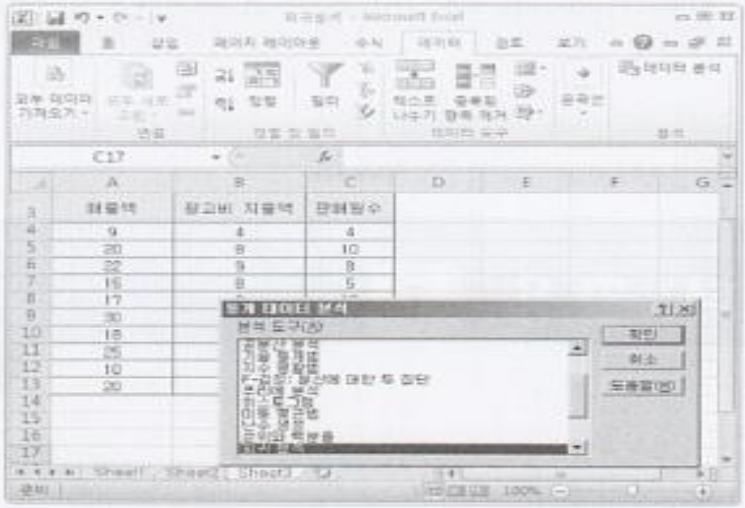


그림 14-4 회귀분석 대화상자

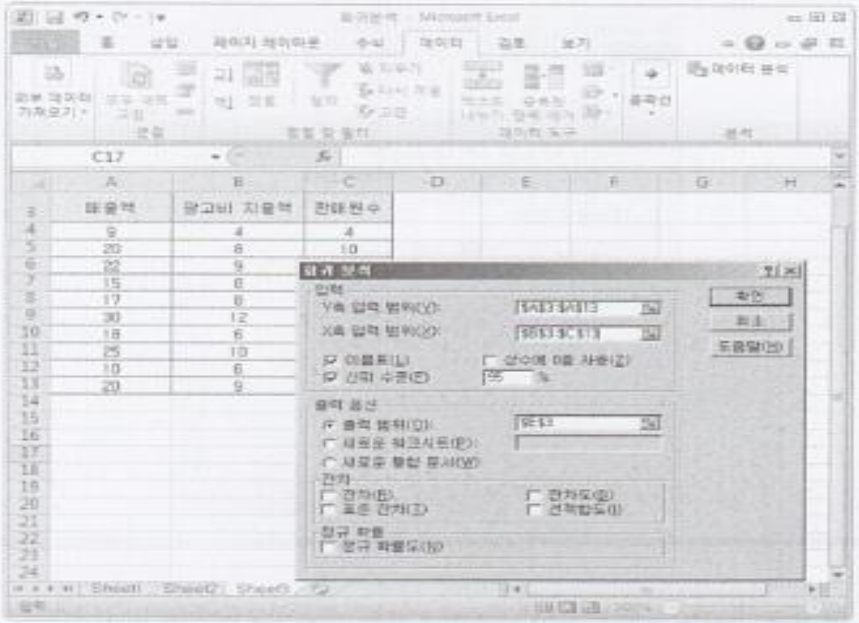


그림 14-4에서 Y축은 매출액자료이고, X축은 광고비 및 판매원 수 자료이므로 해당되는 각 셀의 범위를 대화상자에 입력한다. 이름표에 표시하고 출력범위는 E3셀로 지정한 후 확인을 누르면 그림 14-5와 같이 결과가 주어진다.

그림 14-5 (다중)회귀분석의 결과

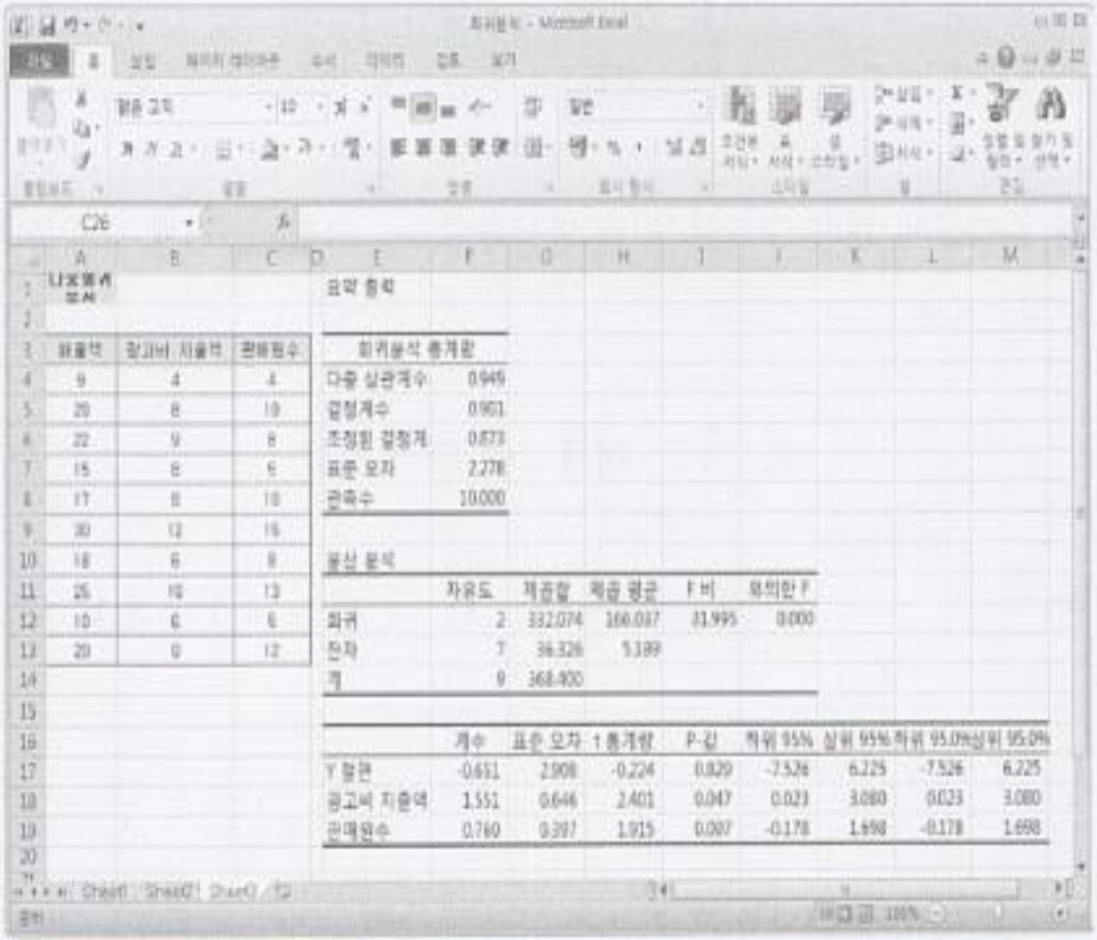


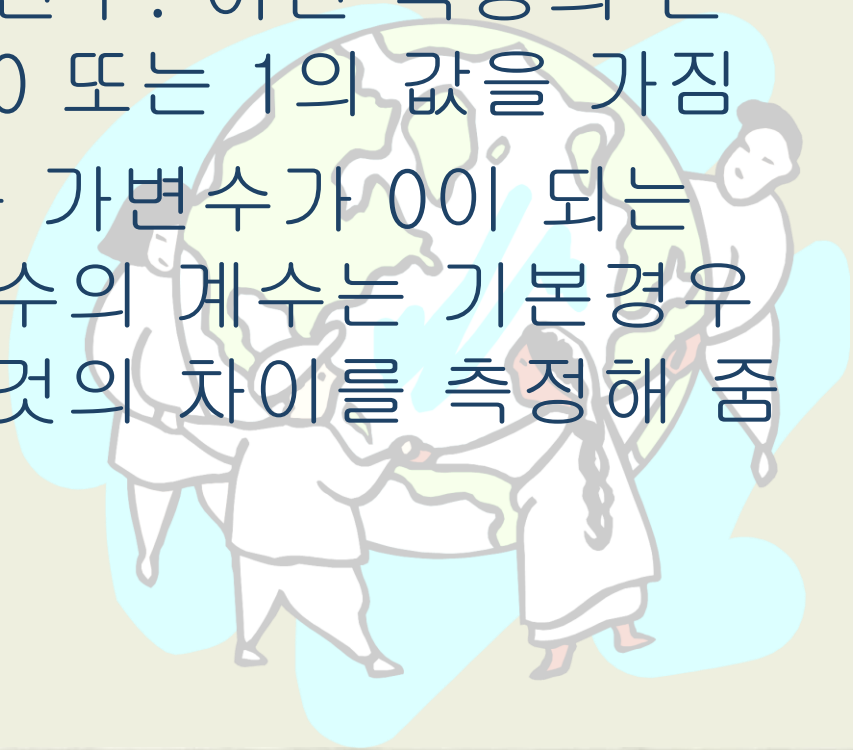
그림 14-5의 결과를 살펴보면 결정계수는 0.901로 주어지는데 이는 종속변수의 변동 중 90.1%가 두 개의 독립변수(여기서는 광고비지출액 및 판매원 수)에 의해 설명됨을 의미한다. 그리고 분산분석의 유의한 F하의 값, 즉 p값이 0.000으로 유의수준 1%보다 낮은 유의수준에서도 추정된 회귀식이 유의함을 알 수 있다. 끝으로 추정된 회귀식은 Y절편이 −0.651이고 독립변수인 광고비지출액의 계수는 1.551, 판매원 수의 계수는 0.760인 $\hat{Y}=-0.651+1.551X_1+0.760X_2$로 표시됨을 알 수 있다. 여기서 광고비지출액의 회귀계수에 대한 p값이 0.05보다 작기 때문에 5% 유의수준에서 매출액에 대한 광고비지출액의 회귀계수가 통계적으로 유의함을 알 수 있다. 그리고 판매원 수의 회귀계수에 대한 p값 또한 0.10보다 작기 때문에 10% 유의수준에서 매출액에 대한 판매원 수의 회귀계수가 유의하다고 할 수 있다.

예제 14-3

예제 1을 참조하여 R^2과 $\overline{R}^2$을 계산하시오.

회귀모형에서의 가변수

- 가변수: 어떤 특성의 존재여부를 나타내기 위해 특별히 구성된 변수. 어떤 특성의 존재여부에 따라 이는 0 또는 1의 값을 가짐
- 가변수의 기본경우는 가변수가 0이 되는 관찰치를 말함. 가변수의 계수는 기본경우와 기본경우가 아닌 것의 차이를 측정해 줌



예제 14-7

소형차의 재판매가격(단위: 천 원)은 차의 연수와 에어컨의 유무에 달려 있다고 한다. Y를 재판매가격, x_1을 차의 연수라 하고 x_2를 에어컨이 있으면 1, 없으면 0이라 하자.

표 14-3의 자료는 20대의 중고차 표본을 보여 준다. 이를 가지고 다중회귀모형을 추정하시오.

표 14-3 중고차에 대한 자료

자동차	재판매가격 (단위: 천 원)	연 수	가변수 (에어컨 유무)
1	4,000	1	0
2	3,050	2	0
3	4,350	1	1
4	3,900	1	0
5	1,950	3	0
6	3,000	2	0
7	1,400	4	1
8	4,500	1	1
9	2,950	2	0
10	900	4	0
11	3,600	2	1
12	4,100	1	0
13	2,100	3	0
14	1,000	4	0
15	2,400	3	1
16	4,000	1	0
17	4,400	1	1
18	2,900	2	0
19	4,450	1	1
20	2,050	3	0

회귀모형의 기본가정에 대한 검정

- 오차항의 독립성 가정
  - 더빈-왓슨 검정
- 정규분포의 가정
- 등분산성(오차항이 독립변수와 관련이 없음)의 가정
  - 골드펠드-콴트 검정

그림 14-13 등분산잔차를 보여주는 (X_i, e_i)의 도표

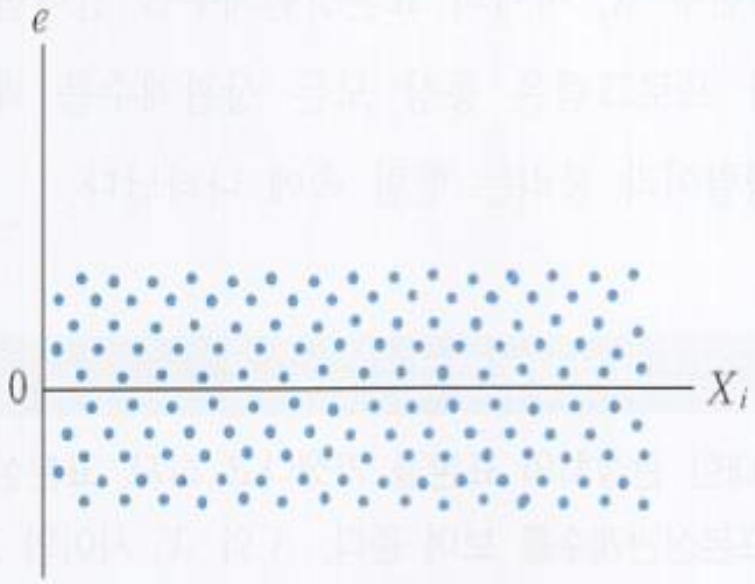
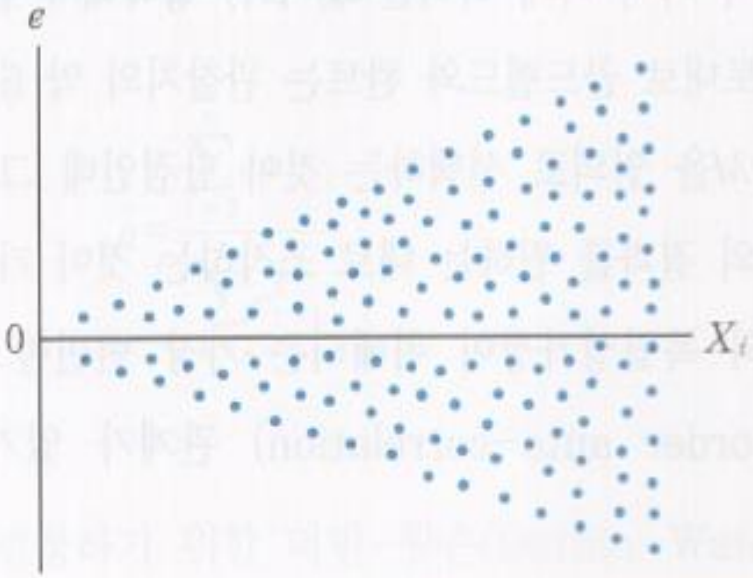


그림 14-14 이분산잔차를 보여주는 (X_i, e_i)의 도표



더빈-왓슨 검정

모회귀모형

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

를 고려하자. 여기서 $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$이고, u_t는 기본가정을 따른다.

$$H_0 : \rho = 0$$
$$H_A : \rho > 0$$

검정통계량: $d = \dfrac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$

기각역: $d < d_{\alpha,L}$이면 H_0을 기각한다.
$d > d_{\alpha,U}$이면 H_0을 채택한다.
$d_{\alpha,L} < d < d_{\alpha,U}$이면 결론을 내릴 수 없다.

하한과 상한 $d_{\alpha,L}$과 $d_{\alpha,U}$는 표본크기 n, 회귀방정식에 포함된 설명변수의 수 k와 검정의 유의수준 α에 의존한다($d_{\alpha,L}$과 $d_{\alpha,U}$의 값을 보여 주는 표는 부록의 표 8에 있음).

때때로 음의 자기상관의 대립가설 즉, $H_1 : \rho < 0$에 대해 검정할 때가 있다. 검정은 검정통계량을 $(4-d)$로 한다는 것을 제외하고는 양의 자기상관에 대한 것과 동일하다.

오차의 등분산성에 대한 골드펠드-콴트 검정

H_0: $\sigma_1^2=\sigma_2^2=\cdots=\sigma_n^2=\sigma^2$

H_A: 오차항은 변수 X_i와 연관이 있다.

이 검정을 수행하기 위해 표본관찰치를 변수 X_i의 크기에 따라서 배열한 후 n_1, M, n_2개의 관찰치로 구성된 세 개의 부분으로 나눈다.

n_1은 첫 번째 부분에 있는 관찰치 수를 나타내고, M은 두 번째 부분에 있는 관찰치 수를 나타내며, n_2는 세 번째 부분에 있는 관찰치 수를 나타낸다고 하자.

일반적으로 $n_1=n_2$이고 M은 훨씬 작게 잡는다. M개 관찰치의 부분은 이 분석에서 생략된다.

한 방정식에서는 첫 번째 n_1개의 관찰치만을 사용하고, 다른 방정식에서는 마지막 n_2개의 관찰치만을 사용하여 두 개의 회귀모형을 추정한다. 첫 번째 회귀방정식으로부터 추정된 분산을 S_1^2으로 나타내고, 두 번째 회귀방정식으로부터의 분산은 S_2^2으로 나타낸다.

F통계량 $F=\frac{S_2^2}{S_1^2}$을 계산한다. 만약 H_0이 사실이면 S_2^2와 S_1^2는 거의 동일할 것이고, 비율 $\frac{S_2^2}{S_1^2}$은 거의 1이 된다. 만약 H_0이 거짓이라면 S_2^2이 S_1^2을 초과하기 때문에 F통계량은 커지게 된다.

이 검정을 수행하면서 $F>F_{\nu_1,\nu_2,\alpha}$이면 H_0을 기각한다. 여기서 $F_{\nu_1,\nu_2,\alpha}$는 $P(F>F_{\nu_1,\nu_2,\alpha})=\alpha$인 F분포의 기각치이다. H_0이 사실일 때, F통계량은 분자자유도 $\nu_1=(n_1-k-1)$과 분모자유도 $\nu_2=(n_2-k-1)$인 F분포를 따른다. 여기서 k는 상수항을 포함하지 않은 설명변수의 수를 나타낸다.