

12장 상관 및 회귀분석

12.1 상관분석

- 두 변량의 관련성을 관찰할 수 있는 쉬운 방법은 한 변량의 값을 X 축으로 하고 다른 변량의 값을 Y 축으로 하여 산점도(scatterplot)를 그려보는 것이다. 두 변량이 관련이 있다면 자료들은 일정한 규칙을 가지고 모일 것이고, 관련이 없다면 흩어져 있을 것이다. **상관분석**(correlation analysis)은 변량들 간의 선형관계의 정도를 분석하는 방법이다. 이를테면, 한 변량이 증가할 때 다른 변량이 얼마나 선형적으로 증가 또는 감소하는가를 조사하는 것이다.
- 두 변량간의 관계의 정도를 구체적인 수치로 나타내어주는 측도를 함께 이용하면 두 변량간의 관계를 더욱 정확하고 객관적으로 파악할 수 있다. 두 변량 사이의 상호관계를 나타내는 측도로서 먼저 공분산(covariance)이 있는데 두 변량 X 와 Y 의 모집단에 대한 공분산은 $Cov(X, Y)$ 로 나타낸다. 두 변량에 대한 n 개의 확률표본 $(X_1, Y_1), \dots, (X_n, Y_n)$ 이 주어졌을 때 이들 표본을 이용한 추정치, 즉, 표본공분산 s_{XY} 는 다음과 같이 정의된다.

$$\begin{aligned} s_{XY} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \frac{1}{n-1} (\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}) \end{aligned}$$

위의 식에서 \bar{X} 와 \bar{Y} 는 각각 X 와 Y 의 표본평균을 나타낸다. 공분산의 의미를 파악하기 위해 X 가 증가할 때 Y 역시 증가하는 경우를 생각해보자. 이 경우 \bar{X} 보다 큰 X 값에 대응하는 Y 값 역시 \bar{Y} 보다 크고, \bar{X} 보다 작은 X 값에 대응하는 Y 값 역시 \bar{Y} 보다 작아진다. 그래서 $(X - \bar{X})(Y - \bar{Y})$ 는 항상 양수값을 가지며 이들의 평균인 공분산 역시 양수 값을 가진다. 이와는 반대로 한 변량의 값이 증가할 때 다른 변량의 값이 감소하면 공분산의 값은 음수가 된다는 것을 쉽게 알 수 있다. 그러므로 공분산을 계산함으로써 양의 상관(즉, 한 변량의 값이 증가하면 다른 변량의 값도 증가)이나 음의 상관(즉, 한 변량의 값이 증가하면 다른 변량의 값은 감소)의 두 변량간 상호관계를 알 수 있다.

- 공분산 자체도 좋은 측도이나 공분산은 X 와 Y 의 단위에 의존하기 때문에 값의 크기에 따른 해석이 어렵고 또한 다른 자료와 비교할 때 불편하다는 단점이 있다. 변량의 종류나 특정단위에 관계없는 측도를 구하기 위해 공분산을 X 와 Y 의 표준편차인 σ_X 와 σ_Y 의 곱으로 나누어 표준화시키는데 이를 모집단상관계수(population correlation coefficient)라 부르고 ρ (‘로오’라고 읽음)라고 표시한다.

$$\text{모집단상관계수: } \rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

- 상관계수 ρ 는 다음과 같이 해석된다.

- 1) ρ 는 -1과 +1 사이의 값을 가지며, ρ 의 값이 +1에 가까울수록 강한 양의 선형 관계를, -1에 가까울수록 강한 음의 상관관계를 나타내며, ρ 의 값이 0에 가까울수록 선형관계는 약해진다.
 - 2) X 와 Y 의 대응되는 모든 값들이 한 직선 상에 위치하면 ρ 의 값은 -1(직선의 기울기가 음인 경우)이나 +1(직선의 기울기가 양인 경우)의 값을 가진다.
 - 3) 상관계수 ρ 는 단지 두 변량의 선형관계만을 나타내는 척도이다. 그러므로, $\rho=0$ 인 경우에 두 변량의 선형상관관계는 없지만 다른 관계는 가질 수 있다.
- 두 변량에 대한 n 개의 표본을 이용한 모집단 상관계수 ρ 의 추정치를 표본상관계수(sample correlation coefficient)라 부르며 r 로 나타낸다. 표본상관계수의 공식은 모집단상관계수의 공식에서 각각의 모수를 추정량으로 대체하여 얻을 수 있다.

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- 표본상관계수 r 은 모집단상관계수 ρ 의 가설검정에도 이용된다. ρ 에 대한 가설검정에서 주로 관심 있는 것은 $H_0: \rho=0$, 즉, 선형상관관계의 존재여부에 대한 것인데, 이에 대한 가설검정은 t 분포를 이용하여 다음과 같이 할 수 있다.

상관계수 ρ 의 검정:

귀무가설: $H_0: \rho=0$

검정통계량: $t_0 = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$

t_0 는 귀무가설 하에서 자유도 $(n-2)$ 인 t 분포를 따른다.

H_0 기각역: 1) $H_1: \rho < 0$ 일 때, $t_0 < -t_{n-2, \alpha}$

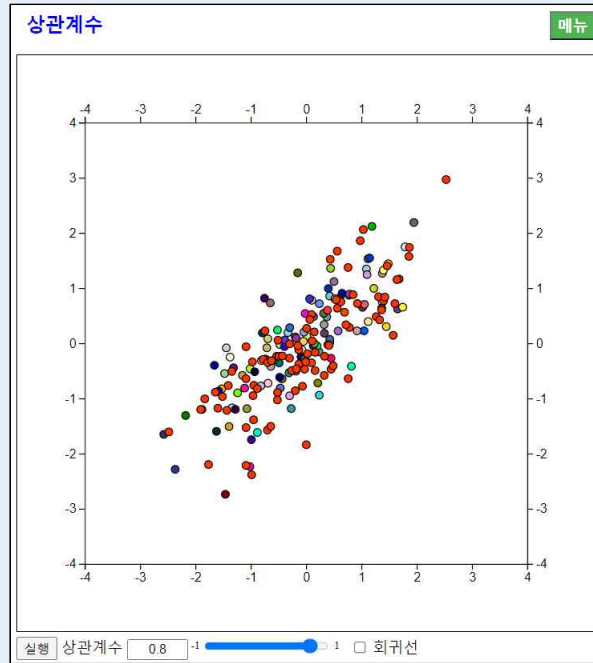
2) $H_1: \rho > 0$ 일 때, $t_0 > t_{n-2, \alpha}$

3) $H_1: \rho \neq 0$ 일 때, $|t_0| > t_{n-2, \alpha/2}$

[예 12.1] 『eStatU』 을 이용해서 여러 상관계수에 대한 산점도 모양을 관찰하여 보자.

풀이

『eStatU』 주메뉴에서 ‘상관계수’를 선택하면 다음과 같은 화면이 나타난다. 여기에서 상관계수를 입력하고 [실행] 버튼을 누르면 이 상관계수를 가는 두 변수에 대한 산점도가 그려진다.



[그림 12.1] 『eStatU』 상관계수 실행

[예 12.2] 동일한 제품을 만드는 10개의 회사들에 대한 광고비와 판매액을 조사한 결과 표 12.1의 자료를 얻었다(단위: 백만원).

표 12.1 판매액과 광고비자료

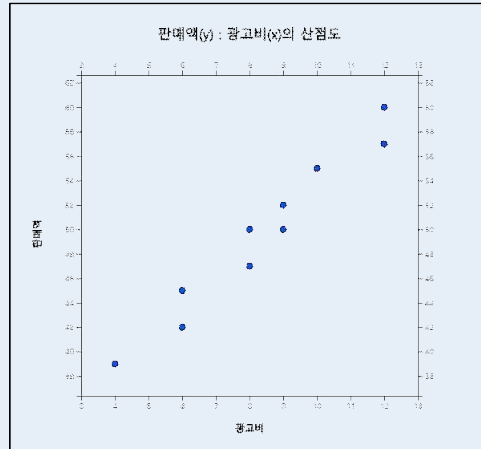
회 사	1	2	3	4	5	6	7	8	9	10
광고비 (X)	4	6	6	8	8	9	9	10	12	12
판매액 (Y)	39	42	45	47	50	50	52	55	57	60

- 1) 『eStat』 을 이용하여 이 자료의 산점도를 작성하고 두 변량의 관련성을 조사하라.
- 2) 공분산과 상관계수를 구하라.
- 3) 광고비와 판매액의 상관계수가 0이라는 가설을 유의수준 0.05에서 검정하라.

1) 『eStat』 을 이용하여 [그림 12.2]와 같이 데이터를 입력한다. 주메뉴의 산점도 아이콘을 누르면 나타나는 변량선택창에서 ‘Y변량’을 판매액, ‘by X변량’을 광고비를 선택하면 [그림 12.3]과 같은 산점도가 나타난다. 산점도는 광고비의 투자가 많을수록 판매액이 증가하는 것을 보여주고, 그뿐만 아니라 증가의 형태가 선형이라는 것을 알 수 있다. .

광고비	판매액	V3	V4
4	39		
6	42		
6	45		
8	47		
8	50		
9	50		
9	52		
10	55		
12	57		
12	60		

[그림 12.2] 『eStat』 데이터 입력



[그림 12.3] 광고비와 판매액에 대한 산점도

2) 공분산과 상관계수를 구하기 위해서는 다음과 같은 표를 만드는 것이 편리하다. 이 표는 회귀분석에서의 계산에도 그대로 이용할 수 있다.

표 12.2 공분산 계산에 필요한 유용한 계산 표

	<i>X</i>	<i>Y</i>	<i>X</i> ²	<i>Y</i> ²	<i>XY</i>
1	4	39	16	1521	156
2	6	42	36	1764	252
3	6	45	36	2025	270
4	8	47	64	2209	376
5	8	50	64	2500	400
6	9	50	81	2500	450
7	9	52	81	2704	468
8	10	55	100	3025	550
9	12	57	144	3249	684
10	12	60	144	3600	720
합계	84	497	766	25097	4326
평균	8.4	49.7			

따라서 공분산과 상관계수를 계산하는데 필요한 항은 다음과 같다.

$$SXX = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = 766 - 10 \times 8.4^2 = 60.4$$

$$SYY = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 = 25097 - 10 \times 49.7^2 = 396.1$$

$$SXY = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} = 4326 - 10 \times 8.4 \times 49.7 = 151.2$$

여기서 *SXX*, *SYY*, *SXY*는 각각 *X*제곱합, *Y*제곱합, *XY*제곱의 합을 나타낸다. 그러므로 공분산과 상관계수는 다음과 같다.

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{151.2}{10-1} = 16.8$$


$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{151.2}{\sqrt{60.4 \times 396.1}} = 0.978$$

이러한 상관계수 값은 자료의 산점도가 두 변량의 강한 양의 상관을 보여주는 것과 일치한다.

3) 검정통계량 t 의 값은 다음과 같다

$$t = \frac{\sqrt{10-2} \cdot 0.978}{\sqrt{1-0.978^2}} = 13.26$$

$t_{80.025} = 2.306$ 보다 크게 되어 가설 $H_0: \rho=0$ 을 기각할 수 있다.

[그림 12.2]와 같이 『eStat』의 변량이 선택된 상태에서 주메뉴의 회귀분석 아이콘 을 누르면 산점도와 회귀선이 나타나고 이 그래프 밑의 ‘상관 및 회귀분석’ 버튼을 누르면 결과저장창에 상관분석([그림 12.5])과 회귀분석 결과가 나타난다. t 값의 결과가 약간 다른데 이는 소숫점 아래 자리수에 관련된 오차이다. 상관분석 검정에 대한 p -값은 0.0001로 유의수준 0.05보다 작으므로 귀무가설을 기각한다는 같은 결론을 얻는다.

회귀분석			
회귀선	y = 28.672 + 2.503 x		
상관계수	r = 0.978	H ₀ : ρ = 0 H ₁ : ρ ≠ 0	t 값 = 13.117 p 값 < 0.0001
결정계수	r ² = 0.956		
추정오차	s = 1.483		

[그림 12.4] 『eStat』의 상관분석 결과



『eStatU』를 이용하여 공분산과 상관계수를 구할 수 있다. 주메뉴에서 ‘산점도-상관분석’을 선택하면 [그림 12.5]와 같은 화면이 나타난다. 여긴 광고비와 판매액을 입력하고, 주제목을 입력한 후 [실행] 버튼을 누르면 공분산과 상관계수의 검정결과와 산점도가 그려진다.

메뉴

산점도 - 상관분석

X 자료 입력

Y 자료 입력

주 제목

세로축 제목 가로축 제목

자료수	n_x	10	n_y	10		
평균	\bar{X}	8.40	\bar{Y}	49.70		
표본분산(n-1)	S_x^2	6.71	S_y^2	44.01	표본 공분산	S_{xy} 16.80
표본표준편차	S_x	2.59	S_y	6.63	표본 상관계수	r 0.978

[가설] $H_0: \rho = 0$ $H_1: \rho \neq 0$ $H_1: \rho > 0$ $H_1: \rho < 0$

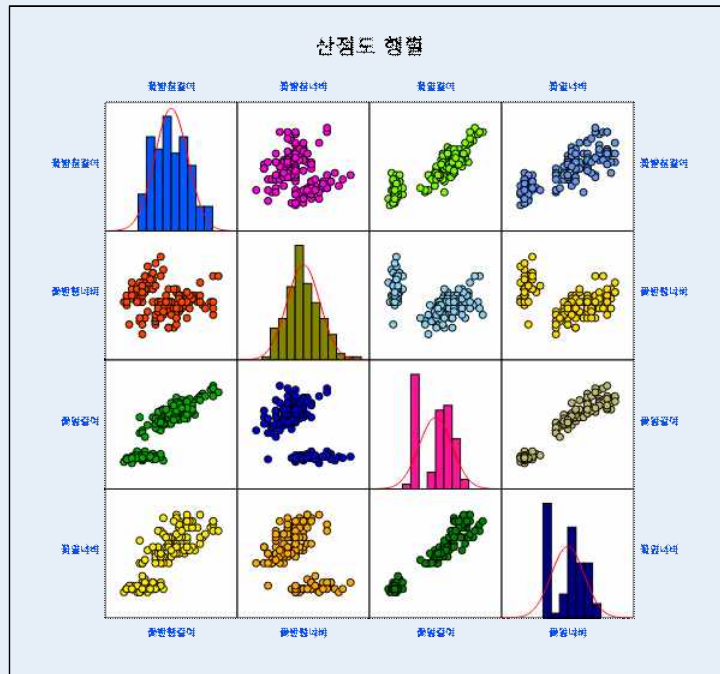
 [검정통계량] $t_0 = \sqrt{(n-2)} r / \sqrt{1-r^2} = 13.117$ p-값 = 0.000

[그림 12.5] 『eStatU』 상관분석

- 분석에 포함된 변량들이 3개 이상인 경우에도 각 두 변량들의 조합에 대해 위의 산점도를 이용하여 그 관계를 볼 수 있고, 상관계수값을 구할 수 있다. 그렇지만 변량들의 관계를 한 눈에 알기 쉽게 하기 위하여 각 변량들 간의 상관계수를 행렬형태로 정리할 수 있는데 이를 **상관계수행렬**(correlation matrix)이라 한다. 『eStat』에서는 상관계수행렬과 그 값들에 대한 유의성 검정의 결과를 함께 보여 준다. 검정의 결과는 t 값과 p -값을 보여준다.

[예 12.3] 『eStat』의 Ex ⇒ 01Korean ⇒ 123중회귀_붓꽃iris.csv 데이터에서 꽃받침길이, 꽃받침너비, 꽃잎길이, 꽃잎너비의 네 변량에 대한 산점도행렬과 상관계수행렬을 구하고 검정결과를 살펴보자.

『eStat』에서 Ex ⇒ 01Korean ⇒ 123중회귀_붓꽃iris.csv 데이터를 불러온 후 ‘회귀분석’ 아이콘을 누른다. 변량선택박스가 나타나면 마우스로 차례로 꽃받침길이, 꽃받침너비, 꽃잎길이, 꽃잎너비의 네 변량을 선택하면 [그림 12.6]과 같은 산점도행렬이 나타난다. 꽃받침길이와 꽃잎길이, 꽃잎너비, 그리고 꽃잎길이와 꽃잎너비가 관련이 있음이 관찰된다.



[그림 12.6] 『eStat』의 산점도행렬

그래프 밑의 선택사항에서 ‘상관 및 회귀분석’을 선택하면 결과저장창에 [그림 12.7]과 같은 데이터의 기초통계량과 상관계수행렬이 검정결과와 함께 나타난다. 꽃받침길이와 꽃받침너비 사이의 상관계수를 제외한 모든 상관계수가 유의함을 알 수 있다.

상관계수행렬					
상관계수 $H_0: \rho=0$ $\rho \neq 0$ t -값 p -값	변량명	변량 1	변량 2	변량 3	변량 4
변량 1	꽃받침길이	1	-0.118 t -값 = -1.440 p -값 = 0.1519	0.872 t -값 = 21.646 p -값 < 0.0001	0.818 t -값 = 17.296 p -값 < 0.0001
변량 2	꽃받침너비	-0.118 t -값 = -1.440 p -값 = 0.1519	1	-0.428 t -값 = -5.768 p -값 < 0.0001	-0.366 t -값 = -4.786 p -값 < 0.0001
변량 3	꽃잎길이	0.872 t -값 = 21.646 p -값 < 0.0001	-0.428 t -값 = -5.768 p -값 < 0.0001	1	0.963 t -값 = 43.387 p -값 < 0.0001
변량 4	꽃잎너비	0.818 t -값 = 17.296 p -값 < 0.0001	-0.366 t -값 = -4.786 p -값 < 0.0001	0.963 t -값 = 43.387 p -값 < 0.0001	1

[그림 12.7] 『eStat』의 여러 변량에 대한 상관계수행렬

12.2 단순선형회귀분석

- **회귀분석**(regression analysis)은 먼저 변량들 간의 관계를 나타내는 타당한 수학적 모형을 설정하고, 변량들의 측정된 값을 이용하여 그 모형을 추정한 다음, 추정한 모형에 의해 변량들 간의 관계를 설명하든지 또는 예측 등의 분석에 응용하는 통계적 방법이다. 예를 들어, 판매액(Y)과 광고비(X)의 관계에 대한 수학적 모형 $Y=f(X)$ 을 설정하였다면 판매액과 광고비와의 관계를 설명할 수 있을 뿐 아니라, 일정한 광고비를 투자했을 때의 판매액을 예측할 수 있을 것이다.
- 이와 같이 회귀분석은 변량들 간의 관련성 정도와 관련형태 조사 및 예측에 그 목적이 있다고 할 수 있다. 회귀분석에서 변량들 간의 관계를 나타내는 수학적 모형을 **회귀식**(regression equation)이라 하며, 서로 관계를 가지고 있는 변량들 중에서 다른 변량에 의해 영향을 받는 변량을 **종속변량**(dependent variable)이라 한다. 종속변량은 우리가 설명하고자 하는 변량으로, 주로 다른 변량들에 대한 반응으로 관측되는 변량이므로 **반응변량**(response variable)이라고도 한다. 그리고 종속변량에 영향을 주는 변량을 **독립변량**(independent variable)이라 부른다. 이는 종속변량을 설명하는데 이용하는 변량이므로 **설명변량**(explanatory variable)이라고도 한다. 앞의 예에서 광고비의 증감에 따른 판매액의 변화를 분석하는 것이 목적이라면 판매액은 종속변량에, 광고비는 독립변량에 해당된다. 그리고 회귀식에 포함된 독립변량의 개수에 따라 단순선형회귀(독립변량이 1개)와 중선형회귀(독립변량이 2개 이상)로 구분하고 있다.

단순선형회귀모형

- 단순선형회귀분석(simple linear regression analysis)은 1개의 독립변량만을 다루며 그 회귀식은 다음과 같이 나타내어진다.

$$Y=f(X;\alpha,\beta)=\alpha+\beta X$$

즉, 회귀식은 독립변량 X 의 일차방정식으로 나타내어지며, α 와 β 는 각각 절편과 기울기를 나타내는 미지의 모수로 **회귀계수**(regression coefficient)라고 한다. 위의 식은 Y 와 X 의 미지의 직선관계를 나타내므로 이를 모집단 회귀식이라 부른다.

- 회귀계수 α 와 β 를 추정하기 위해서는 종속변량 Y 와 독립변량 X 의 관측값들, 즉, 표본이 필요하다. 이때 이들 관측값들은 일반적으로 모두 일직선상에 위치하지는 않는다. 왜냐하면, Y 와 X 가 정확한 선형관계를 가지고 있다 하더라도 관측값에는 측정오차가 있을 수도 있고, 또는 실제로 Y 와 X 가 정확하게 선형관계를 형성하지 않을 수도 있기 때문이다. 그러므로 이들 오차를 함께 생각하여 회귀식을 다음과 같이 적을 수 있다.

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad i=1,2,\dots,n$$

여기에서 i 는 전체 n 개의 관측값 중 i 번째 값을 나타내는 첨자이고, ϵ_i 들은 평균

이 0, 분산이 σ^2 인 서로 독립인 오차를 나타내는 확률변량으로 관측값 Y_i 가 모집단 회귀식으로부터 ε_i 만큼 떨어져 있음을 의미한다. 위의 식은 미지의 모수 α , β 와 σ^2 을 포함하므로 이를 모집단 회귀모형이라 부른다.

- 이에 대해 표본을 이용하여 추정된 회귀계수를 a 와 b 로 나타내면 **적합된(fitted)** 회귀식을 다음과 같이 적을 수 있고 이를 표본 회귀식이라 한다.

$$\widehat{Y}_i = a + bX_i$$

이 식에서 \widehat{Y}_i 은 적합된 회귀식에 의해 X_i 에서 예측된 Y 의 값을 나타낸다. 이들 예측된 값들은 Y 의 실제 관측된 값들과 일치할 수는 없는데 이 두 값의 차이를 **잔차(residual)**라 부르고 e_i 로 표시한다.

$$\text{잔차} : e_i = Y_i - \widehat{Y}_i, \quad i=1,2,\dots,n$$

- 회귀분석에서는 관측할 수 없는 오차인 ε_i 에 대해 몇 가지 가정을 하는데, 표본 값을 이용하여 계산되는 잔차 e_i 는 ε_i 와 비슷한 성질을 가지므로 이들 가정의 타당성을 조사하는데 중요하게 사용된다. (잔차분석 참조.)

회귀계수의 추정

- 표본 $(X_1, Y_1), \dots, (X_n, Y_n)$ 이 주어졌을 때 이를 대표하는 직선은 여러 가지로 그어질 수 있다. 회귀분석의 주된 목적 중의 하나가 예측이므로, 우리는 추정된 회귀식을 이용하여 Y 의 값을 예측할 때 발생하는 오차인 잔차들을 가장 작게 하여 줄 수 있는 식을 선택하고자 한다. 그러나 모든 점에서 잔차의 값을 최소화할 수는 없고 잔차의 크기를 “전체적”으로 작게 하는 방법을 선택하여야 한다. 이러한 방법들 중 가장 널리 사용되는 것은 잔차의 제곱들의 합을 최소로 하는 회귀식을 구하는 방법으로 이를 **최소제곱법(method of least squares)**이라 한다.

최소제곱법

각 관측값에서 발생하는 오차들의 제곱합이 최소가 되도록 회귀계수를 추정하는 방법. 즉,

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2$$

가 최소가 되는 α 와 β 의 값을 구한다.

- 최소제곱법에 의해 α 와 β 의 값을 구하기 위해서는 위의 제곱합을 α 와 β 에 대해 각각 편미분하여 영으로 놓고 α 와 β 에 대해 풀면 된다. 이때 얻어지는 α 와 β 의 값을 각각 a 와 b 로 나타내면, 구하고자 하는 두 값은 아래의 두 식을 만족하게 된다.

$$an + b\sum_{i=1}^n X_i = \sum_{i=1}^n Y_i$$

$$a\sum_{i=1}^n X_i + \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

위의 식을 **정규방정식**(normal equation)이라 한다. 이 정규방정식의 해를 α 와 β 의 **최소제곱추정량**(least squares estimator)이라 하며 다음과 같이 주어진다.

최소제곱추정량 :

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

여기서 b 의 분모와 분자를 각각 $n-1$ 로 나누어 주면 $b = s_{XY}/s_X^2$ 로 쓸 수 있고, 상관계수는 $r = \frac{s_{XY}}{s_X s_Y}$ 이므로 $s_{XY} = r s_X s_Y$ 이다. 따라서 상관계수를 알면 기울기는

$$b = \frac{s_{XY}}{s_X^2} = \frac{r s_X s_Y}{s_X^2} = r \frac{s_Y}{s_X}$$

로도 계산할 수 있다.

회귀직선의 적합도

- 가정된 회귀직선을 추정한 다음에는 그 회귀식이 얼마나 타당한가를 조사하여야 한다. 이를테면, 회귀분석의 목적은 종속변량을 독립변량의 함수로 설명하고자 함이므로 과연 그 설명의 정도가 어느 정도인지를 알아 볼 필요가 있다. 이와 같은 타당성 조사에는 **잔차표준오차**(residual standard error)와 **결정계수**(coefficient of determination)가 사용된다.
- 잔차표준오차 s 는 관측값들이 추정회귀직선의 주위에 흩어져 있는 정도를 나타내는 척도이다. 먼저 잔차들의 표본분산을 다음과 같이 정의할 수 있는데

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

잔차표준오차 s 는 s^2 의 제곱근으로 정의된다. 그리고 s^2 은 Y 의 값들이 모집단 회귀직선을 중심으로 퍼져있는 정도를 나타내는 σ^2 의 추정량이 된다. s 또는 s^2 의 값이 작으면 관측값들이 추정회귀직선에 근접해 있음을 나타내고, 이는 역으로 추정회귀직선이 두 변량간의 관계를 잘 대표한다고 이야기할 수 있다.

- 그러나 잔차표준오차 s 는 그 값이 “작으면” 좋은 것이지만 어느 정도의 값이 작은 것인지는 분명하지가 않다. 또한 s 의 값의 크기는 Y 의 단위에 의존한다. 이러한 단점을 없애기 위해서는 상대적인 척도가 필요한데, 여기에서 정의할 결정계

수는 Y_i 들이 가지는 총변량 중 회귀직선에 의해 설명되는 변량의 비(ratio)로서 주어지므로 변량의 종류와 단위에 관계없이 사용할 수 있는 상대적 측도이다.

- 9장의 분산분석에서와 같이 회귀분석에서도 다음과 같은 제곱합과 자유도의 분할이 성립한다.

제곱합과 자유도의 분할 :

$$\text{제곱합: } SST = SSE + SSR \quad (12-13)$$

$$\text{자유도: } (n-1) = (n-2) + 1 \quad (12-14)$$

- 위의 세 가지 제곱합에 대한 설명은 아래와 같다.

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2 :$$

Y_i 의 관측값들이 가지는 총변동을 나타내는 제곱합으로 이를 총제곱합(total sum of squares, SST)이라 한다. 이 SST 는 자유도 $(n-1)$ 를 가지며 이 자유도로 나누면 Y_i 값들의 표본분산이 된다.

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

잔차들의 제곱합으로 Y_i 의 총변동 중 설명 안된 변동(unexplained variation)을 나타내며 이를 오차제곱합(error sum of squares, SSE)이라 한다. 이 제곱합의 계산을 위해서는 두 개의 모수 α 와 β 를 추정해야 하므로, SSE 는 $(n-2)$ 의 자유도를 가진다. 잔차들의 표본분산 s^2 의 계산에서 $(n-2)$ 로 나누어주는 이유는 여기에 있다.

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 :$$

Y_i 의 총변동 중 회귀식에 의해 설명된 변동(explained variation)을 나타내며 이를 회귀제곱합(regression sum of squares, SSR)이라 한다. 이 제곱합은 자유도 1을 가진다.

- 만일 추정된 회귀식이 모든 표본의 변동을 완전히 설명하고 있다면 (즉, 모든 관측값들이 표본회귀직선 위에 있을 경우), 설명 안된 변동 SSE 는 0이 될 것이다. 따라서, 총제곱합 SST 중에서 SSE 가 차지하는 부분이 작으면, 또는 SSR 이 차지하는 부분이 크면 추정된 회귀모형의 적합도가 높다고 할 수 있다. 그러므로, 총변동 SST 중에서 설명된 변동 SSR 이 차지하는 비,

$$R^2 = \frac{\text{설명된 변동}}{\text{총변동}} = \frac{SSR}{SST}$$

을 **결정계수**(coefficient of determination)라 정의하고 회귀직선의 적합도를 나타내는 측도로 사용한다. 결정계수의 값은 항상 0 과 1 사이에 있고 그 값이 1에 가까울수록 표본들이 회귀직선 주위에 밀집되어 있음을 뜻하고 이는 추정된 회귀식이 관측값들을 잘 설명하고 있다는 것을 뜻한다.

회귀의 분산분석

- 앞에서 구한 세 가지 제곱합을 자유도로 나누면 각각은 일종의 분산이 된다. 이를테면, SST 를 자유도 $(n-1)$ 로 나누면 Y 의 관측값 Y_1, Y_2, \dots, Y_n 의 표본분산이며, SSE 를 자유도 $(n-2)$ 로 나누면 오차의 분산인 σ^2 의 추정량 s^2 이 된다. 이런 이유로 제곱합의 분할을 이용하여 회귀분석과 관련된 문제를 다루는 것을 회귀의 분산분석이라 한다. 계산된 제곱합과 자유도 등 분산분석에 필요한 정보는 표 12.3과 같은 분산분석표에 정리될 수 있다.

표 12.3 단순 선형회귀의 분산분석표

요인	제곱합	자유도	평균제곱	F값
처리	SSR	1	MSR=SSR/1	Fo=MSR/MSE
오차	SSE	n-2	MSE=SSE/(n-2)	
전체	SST	n-1		

- 제곱합을 자유도로 나눈 값을 **평균제곱**(mean square)이라 하는데 표 12.3에는 **회귀평균제곱**(regression mean square, MSR)과 **오차평균제곱**(error mean square, MSE)이 각각 정의되어 있다. 식에서 알 수 있듯이 MSE 는 σ^2 의 추정량 s^2 과 같은 통계량이다.
- 마지막 열에 주어진 F값은 가설 $H_0:\beta=0$ 대 $H_1:\beta\neq 0$ 의 검정에 사용된다. 만약 β 가 0 이 아니라면 가정된 회귀식이 타당하여 Y 의 변동이 회귀식에 의해 상당 부분 설명될 것이므로 F값이 클 것을 예상할 수 있다. 그러므로, 우리는 역으로, 계산된 F 비가 충분히 크면 β 가 0이 아니라고 결정할 수 있다. 모집단 회귀모형에서 언급된 오차항에 대한 가정이 성립하고 오차항이 정규분포를 따르면 귀무가설 $H_0:\beta=0$ 하에서 F 비는 자유도 1과 $(n-2)$ 의 F 분포를 따름을 보일 수 있다. 그러므로, 만약 $F_0 > F_{1, n-2; \alpha}$ 이면 $H_0:\beta=0$ 을 기각할 수 있다.

단순 선형회귀분석에서의 F 검정:

가설: $H_0:\beta=0, \quad H_1:\beta\neq 0$

검정: $F_0 = \frac{MSR}{MSE} > F_{1, n-2; \alpha}$ 이면 H_0 를 기각

(『eStat』에서는 이 검정에 대한 p -값을 계산하여 주므로 이 p -값을 이용하여 검정한다. 즉, p -값이 유의수준보다 작으면 귀무가설 H_0 을 기각한다.)

[예 12.4] [예 12.2]의 광고비 예제에서

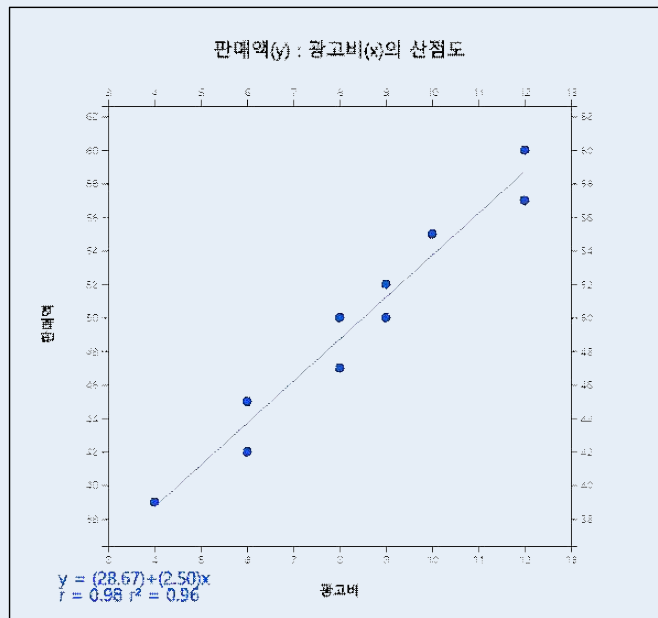
- 1) 판매액을 종속변량, 광고비를 독립변량으로 할 때 절편과 기울기의 최소제곱 추정값을 구하라.
- 2) 광고비가 10만큼 지출하였을 때의 판매액을 예측하라.
- 3) 광고비와 판매액에 관한 자료에서 잔차표준오차와 결정계수의 값을 계산하라.
- 4) 분산분석표를 작성하고 유의수준 5%로 F 검정을 하여보라.

1) [예 12.2]에서 이미 절편과 기울기를 구하는데 필요한 계산을 하였는데 이를 이용해서 절편과 기울기를 구하면 다음과 같다.

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{151.2}{60.4} = 2.503$$

$$a = \bar{Y} - b\bar{X} = 49.7 - 2.503 \times 8.4 = 28.627$$

따라서 적합된 회귀식은 $\hat{Y}_i = 28.672 + 2.5033X_i$ 이다. [그림 12.8]은 원래의 자료 위에 적합된 회귀식을 그려본 것이다. 위에서 2.5033은 직선의 기울기로 광고비가 1만큼, 즉, 백만원씩 증가하면 판매액은 약 2.5백만원씩 증가함을 의미한다.



[그림 12.8] 『eStat』의 단순선형회귀 산점도와 회귀선

2) 광고비가 10인 회사에서의 판매액의 예측은 위의 표본회귀식을 이용하여 $29.672 + (2.5033)(10) = 53.705$ 즉, 약 53.7백만원의 판매액이 예상된다. 이는 광고비가 10백만원인 모든 회사들의 판매액이 53.7백만원이 된다는 것이 아니라 그 회사들의 판매액 평균이 그 정도 된다는 것이다. 그러므로, 개개의 회사에 있어서는 약간의 차이가 있을 수 있다.

3) 잔차표준오차와 결정계수를 구하기 위해서는 다음과 같은 표 12.4를 만들면 편리하다. 여기서 각각의 X_i 값에서 판매액의 추정값은 적합된 회귀식을 이용한다.

$$\hat{Y}_i = 28.672 + 2.5033X_i$$

표 12.4 표준오차와 결정계수를 구하는데 유용한 계산

	X	Y	\hat{Y}_i	SST ($Y_i - \bar{Y}$) ²	SSR ($\hat{Y}_i - \bar{Y}$) ²	SSE ($Y_i - \hat{Y}_i$) ²
1	4	39	38.639	114.49	122.346	0.130
2	6	42	43.645	59.29	36.663	2.706
3	6	45	43.645	22.09	36.663	1.836
4	8	47	48.651	7.29	1.100	2.726
5	8	50	48.651	0.09	1.100	1.820
6	9	50	51.154	0.09	2.114	1.332
7	9	52	51.154	5.29	2.114	0.716
8	10	55	53.657	28.09	15.658	1.804
9	12	57	58.663	53.29	80.335	2.766
10	12	60	58.663	106.09	80.335	1.788
합계	84	497	496.522	396.1	378.429	17.622
평균	8.4	49.7				

표 12.4에서 $SST = 396.1$, $SSR = 378.429$, $SSE = 17.622$ 이다. 여기서 $SST = SSE + SSR$ 인 관계가 정확히 맞지 않는 것은 소수이하 자리수 계산의 오차 때문이다. 잔차들의 표본분산은 다음과 같다.

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{17.622}{(10-2)} = 2.203$$

따라서 잔차표준오차는 $s = 1.484$ 이다. 결정계수는 다음과 같다.

$$R^2 = \frac{SSR}{SST} = \frac{378.429}{396.1} = 0.956$$

이것은 관측된 10개의 판매액이 가지는 총변동의 95.6%를 광고비라는 변량을 사용한 단순선형회귀모형으로 설명할 수 있다는 것을 의미하므로 이 회귀직선은 상당히 유용하다고 할 수 있다.

[그림 12.8] 그래프 밑의 선택사항에서 '상관 및 회귀분석' 버튼을 누르면 [그림 12.9]와 같은 결정계수와 추정오차를 보여준다.

회귀분석				
회귀선	y = 28.672 + 2.503 x			
상관계수	r = 0.978	H ₀ : ρ = 0 H ₁ : ρ ≠ 0	t 값 = 13.117	p 값 < 0.0001
결정계수	r ² = 0.956			
추정오차	s = 1.483			

[그림 12.9] 『eStat』의 결정계수와 추정오차 계산 결과

표 12.4에서 계산한 제곱합을 이용하여 분산분석표를 작성하면 다음과 같다.

요인	제곱합	자유도	평균제곱	F값
처리 오차	378.42	1	$MSR = 378.42/1 = 378.42$	$F_0 = 378.42/2$
전체	17.62	10-2	$MSE = 17.62/8 = 2.20$	
전체	396.04	10-1		

F 검정에서는 계산된 $F_{값} = 172.0$ 은 $F_{1.8;0.05} = 5.32$ 보다 매우 크므로 유의수준 $\alpha = 0.05$ 에서 'β가 0 이다'라는 가설을 기각할 수 있다.
 [그림 12.8] 그래프 밑의 선택사항에서 '상관 및 회귀분석' 버튼을 누르면 [그림 12.10]과 같은 분산분석 결과를 보여준다.

[분산분석]					
요인	제곱합	자유도	평균제곱	F 값	p 값
회귀	378.501	1	378.501	172.052	< 0.0001
오차	17.599	8	2.200		
전체	396.100	9			

[그림 12.10] 『eStat』의 분산분석 결과

회귀분석에서의 추론

- 모집단 회귀모형의 오차항 ϵ 이 평균이 0, 분산이 σ^2 인 정규분포를 따른다는 가정 하에서 회귀계수 α 와 β , 그리고 그 외 모수들에 관한 추론을 할 수가 있다. 참고로 회귀모형 $Y = \alpha + \beta X + \epsilon$ 에서 위의 가정 하에 Y 는 평균이 $\alpha + \beta X$ 이고 분산이 σ^2 인 정규분포를 따름을 알 수 있다.

1) 모수 β에 관한 추론

회귀직선의 기울기인 모수 β는 종속변량과 독립변량간의 선형관계 존재여부와 그 정도를 나타낸다. β에 관한 추론은 아래와 같이 요약될 수 있는데, 특히 가설 $H_0: \beta = 0$ 에 대한 검정은 독립변량이 종속변량을 유의적으로 설명하는지에 대한 것으로 중요하게 사용된다.

모수 β에 관한 추론

점추정량: $b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$, $b \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2})$

점추정량의 표준오차: $SE(b) = \frac{s}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$

신뢰구간: $\beta = b \pm t_{n-2; \alpha/2} \cdot SE(b)$

가설검정:

귀무가설: $H_0: \beta = \beta_0$

검정통계량: $t = \frac{b - \beta_0}{SE(b)}$

H_0 기각역: 대립가설이 $H_1: \beta < \beta_0$ 이면 $t < -t_{n-2; \alpha}$
 대립가설이 $H_1: \beta > \beta_0$ 이면
 대립가설이 $H_1: \beta \neq \beta_0$ 이면 $|t| > t_{n-2; \alpha/2}$

2) 모수 α 에 관한 추론

회귀직선의 절편인 모수 α 에 관한 추론은 아래 표와 같이 요약될 수 있다. 모수 α 는 독립변량 X 가 0일 때의 반응변량 Y 의 평균값을 나타내므로 대부분의 분석에서는 관심의 대상이 되지 않는다.

모수 α 에 관한 추론

점추정량: $a = \bar{Y} - b\bar{X}$, $a \sim M\left(\alpha, \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \cdot \sigma^2\right)$

점추정량의 표준오차: $SE(a) = s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$

신뢰구간: $\alpha = a \pm t_{n-2; \alpha/2} \cdot SE(a)$

가설검정:

 귀무가설: $H_0: \alpha = \alpha_0$

 검정통계량: $t = \frac{a - \alpha_0}{SE(a)}$

H_0 기각역: 대립가설이 $H_1: \alpha < \alpha_0$ 이면 $t < -t_{n-2; \alpha}$
 대립가설이 $H_1: \alpha > \alpha_0$ 이면 $t > t_{n-2; \alpha}$
 대립가설이 $H_1: \alpha \neq \alpha_0$ 이면 $|t| > t_{n-2; \alpha/2}$

3) Y 의 평균값에 관한 추론

X 의 임의의 점 $X=X_0$ 에서 종속변량 Y 는 평균값 $\mu_{Y|X} = \alpha + \beta X_0$ 를 가진다. 이 값을 추정한다는 것은 Y 의 평균값을 예측하는 것과 같은 의미이므로 $\mu_{Y|X}$ 역시 중요한 모수로 취급된다.

평균값 $\mu_{Y|X} = \alpha + \beta X_0$ 에 관한 추론

점추정량: $\hat{Y}_0 = a + bX_0$

점추정량의 표준오차: $SE(\hat{Y}_0) = s \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$

신뢰구간: $\mu_{Y|X} = \hat{Y}_0 \pm t_{n-2; \alpha/2} \cdot SE(\hat{Y}_0)$

평균값 $\mu_{Y|X}$ 의 신뢰구간 공식을 보면 표준오차가 주어진 X 의 값에 의존하므로 신뢰구간의 폭은 주어진 X 의 값에 따라 달라진다. 표준오차의 공식에서 알 수 있듯이 이 폭은 $X = \bar{X}$ 일 때에 가장 좁고 X 가 \bar{X} 에서 멀어질수록 넓어진다. X 의 각 점에서 Y 의 평균값에 대한 신뢰구간을 구한 다음 상한들과 하한들을 제각기 연결하면 표본회귀식 위아래로 곡선을 형성하게 되는데 이를 회귀직선의 신뢰대(confidence band)라 한다.

잔차분석

- 앞절에서의 각 모수에 대한 추론은 모두 모집단 회귀모형에 포함된 오차항 ε 에 대한 몇 가지 가정을 바탕으로 하고 있다. 그러므로, 타당한 추론을 하기 위해서는 이들 가정들의 성립이 중요한 전제조건이 된다. 그러나 오차항 ε 은 관측될 수 없는 값이기 때문에 이들의 일종의 추정량인 잔차를 이용하여 이들 가정의 타당성을 조사하는데 이를 **잔차분석(residual analysis)**이라 한다.
- 먼저 회귀분석에서의 가정을 살펴보자.

회귀분석에서의 가정

- $A1$: 가정된 모형 $Y = \alpha + \beta X + \varepsilon$ 은 옳다.
- $A2$: 오차 ε_i 의 평균값은 0이다.
- $A3$: (등분산성) 모든 ε_i 의 분산은 σ^2 으로 동일하다.
- $A4$: (독립성) 오차 ε_i 들은 서로 독립이다.
- $A5$: (정규성) 오차 ε_i 들은 정규분포를 따른다.

- 이들 가정들의 자세한 의미는 참고문헌을 살펴보기 바란다. 이들 가정들의 타당성은 일반적으로 잔차의 산점도를 이용하여 조사되는데 각각의 가정을 위해 주로 사용되는 산점도는 다음과 같다.

- 1) 잔차 대 예측값 (즉, e_i 대 \hat{Y}_i) : $A3$
- 2) 잔차 대 독립변량 (즉, e_i 대 X_i) : $A1$
- 3) 잔차 대 관측순서 (즉, e_i 대 i) : $A2, A4$

위의 산점도들에서는 잔차들이 0을 중심으로 특정한 경향을 보이지 않고 랜덤하게 나타나면 각 가정이 타당함을 의미한다.

- 오차항 ε 이 정규분포를 따른다는 가정은 자료가 많은 경우 잔차들의 히스토그램을 작성하여 정규분포의 모양과 비슷한지를 보아 그 타당성을 조사할 수 있다. 또 다른 방법은 잔차들의 Q-Q 산점도(quantile-quantile plot)를 이용하는 방법이다. 일반적으로 잔차의 Q-Q 산점도가 직선을 형성하면 정규분포를 따른다고 볼 수 있다.
- 잔차들도 종속변량 Y 의 단위에 의존되므로 잔차분석을 할 때 일관성있는 해석을 위하여 잔차들의 표준화된 값을 사용하는데 이를 표준화 잔차라 한다. 위에서 설명한 잔차들의 산점도와 Q-Q 산점도 모두 표준화 잔차를 사용하여 작성된다. 특히 표준화 잔차의 값이 ± 2 를 벗어나면 이상값 또는 특이값을 의심할 수 있다.

[예 12.5] [예 12.2]의 광고비 예제에서

- 1) 각 모수에 대한 추론을 하라.
- 2) 『eStat』을 이용하여 검정결과와 신뢰대를 확인하라.
- 3) 잔차들의 산점도와 Q-Q 산점도를 작성하라.

1) β 에 관한 추론

β 의 점추정량의 값은 $b=2.5033$ 이고, b 의 표준오차는 다음과 같다.

$$SE(b) = \frac{s}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{1.484}{\sqrt{60.4}} = 0.1908$$

그러므로 β 의 95% 신뢰구간은 $t_{8;0.025}=3.833$ 을 이용하여

$$2.5033 \pm (3.833)(0.1908)$$

$$2.5033 \pm 0.7313$$

즉, (1.7720, 3.2346)으로 얻는다.

$H_0: \beta=0$ 대 $H_1: \beta \neq 0$ 의 가설에 대한 검정을 위한 검정통계량의 값은 다음과 같다.

$$t = \frac{2.5033 - 0}{0.1908} = 13.12$$

$t_{8;0.025}=3.833$ 이므로 유의수준 $\alpha=0.05$ 에서 귀무가설 $H_0: \beta=0$ 은 기각된다.

이 양측검정에 대한 결과는 위의 신뢰구간에서도 얻을 수 있다. 즉, 95% 신뢰구간 (1.7720, 3.2346)은 0을 포함하지 않으므로 ‘ β 가 0 이다’라는 가설을 기각할 수 있다.

α 에 관한 추론

α 의 점추정량의 값은 $a=29.672$ 이고 표준오차는 다음과 같다.

$$SE(a) = s \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}} = 1.484 \sqrt{\frac{1}{10} + \frac{8.4^2}{60.4}} = 1.670$$

t 값은 $29.672/1.67=17.1657$ 이고 $t_{8;0.025}=3.833$ 이므로, 역시 유의수준 $\alpha=0.05$ 에서 ‘절편이 0 이다’는 가설은 기각된다.

Y 의 평균값에 관한 추론

『eStat』에서는 $\mu_{Y|X}$ 의 추정값인 \hat{Y} 의 표준오차를 각 관찰점에서 계산하여 준다. 예를 들어, $X=8$ 에서 점추정값은 $\hat{Y}=49.699$ 이고, 이에 대한 표준오차는 0.475 이다. 그러므로, $\mu_{Y|X}$ 에 대한 95% 신뢰구간은

$$49.699 \pm (3.833)(0.475)$$

$$49.699 \pm 1.821$$

즉, (46.878, 50.520)으로 주어진다. X 의 다른 점에서도 같은 방법으로 신뢰구간을 구할 수 있는데, 이를테면,

$$X=4 \text{에서 } 39.685 \pm (3.833)(0.962) \Rightarrow (33.998, 43.372)$$

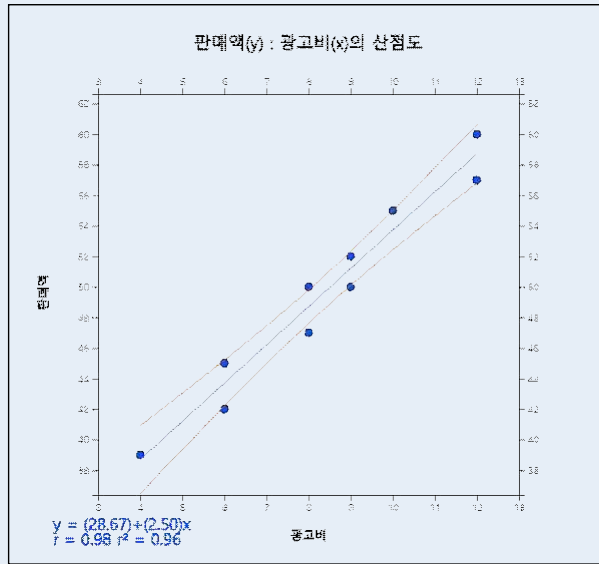
$$X=6 \text{에서 } 43.692 \pm (3.833)(0.656) \Rightarrow (41.178, 46.206)$$

$$X=9 \text{에서 } 51.202 \pm (3.833)(0.483) \Rightarrow (49.351, 53.053)$$

$$X=12 \text{에서 } 59.712 \pm (3.833)(0.832) \Rightarrow (56.063, 61.361)$$

를 얻을 수 있다. 본문에서 언급한 것처럼 신뢰구간의 폭은 X 가 \bar{X} 에서 멀어질수록 넓어짐을 알 수 있다.

2) [그림 12.8] 그래프 밑의 선택사항에서 '신뢰대'를 선택하면 산점도-회귀선 그래프에 [그림 12.11]과 같은 신뢰대를 그려준다. '상관 및 회귀 분석' 버튼을 누르면 결과저장창에 [그림 12.12]와 같은 각 모수의 추론 결과를 보여준다.

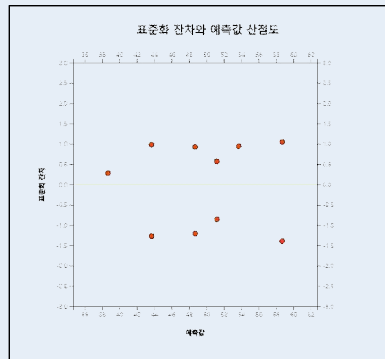


[그림 12.11] 『eStat』의 회귀 신뢰대

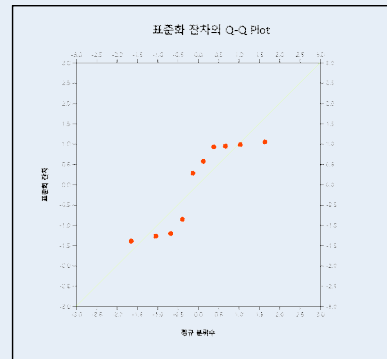
모수	추정값	표준오차	t 값	p 값
절편	28.672	1.670	17.166	< 0.0001
기울기	2.503	0.191	13.117	< 0.0001

[그림 12.12] 『eStat』의 각 회귀모수 추론

[그림 12.8] 그래프 밑의 선택사항에서 [잔차와 예측값] 버튼을 누르면 [그림 12.13]과 같은 표준화 잔차와 예측값 산점도를 그려주고, '잔차 Q-Q 산점도'를 누르면 [그림 12.14]가 나타난다. 잔차의 산점도는 특이 사항이 없으나, Q-Q 산점도는 직선의 형태에서 많이 벗어나므로 오차항의 정규성에는 어느 정도 의심이 간다고 할 수 있다. 이러한 경우에는 반응변량의 값들을 로그변환이나 제곱근변환을 하여 다시 분석할 필요가 있다.



[그림 12.13] 광고비와 판매액 회귀결과의 잔차분석

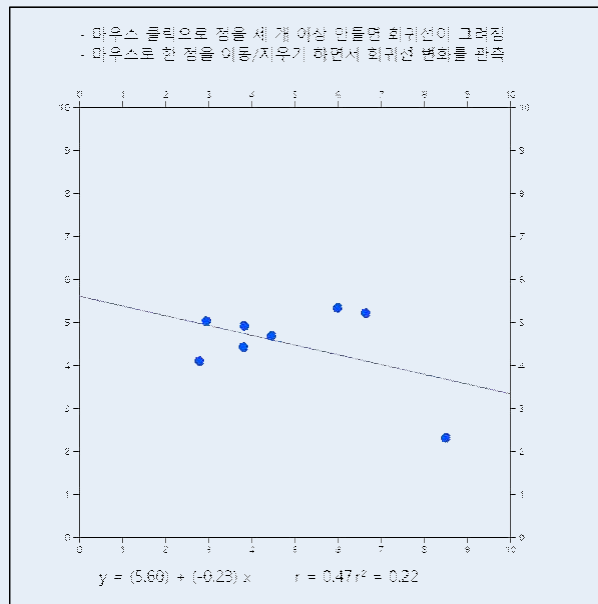


[그림 12.14] 광고비와 판매액 회귀결과의 Q-Q 산점도

- 『eStatU』에서는 회귀선이 극단점에 얼마나 영향을 받는지 실험할 수 있다

[예 12.6] 『eStatU』을 이용해서 선형회귀분석의 극단점의 영향을 실험하여 보라.

『eStatU』주메뉴에서 ‘회귀선 실험’을 선택하면 다음과 같은 화면이 나타난다. 여기에서 화면에 마우스를 클릭하면 점이 찍힌다. 여러 개의 점을 만들면 그때 마다 회귀선이 바뀌는 정도가 얼마나 되는지 살펴볼 수 있다. 만들어진 점은 마우스로 눌러 이동하면서 상관계수와 결정계수가 얼마나 민감한지를 관찰할 수 있다.



[그림 12.15] 『eStatU』 회귀선 실험

12.3 중선형회귀분석

- 회귀분석의 실제 응용에 있어서는 독립변량이 1개인 단순선형회귀보다는 독립변량이 2개 이상 포함된 중회귀모형이 더욱 빈번하게 사용된다. 왜냐하면, 종속변량이 단 한 개의 독립변량만으로 충분하게 설명되는 경우는 드물고, 대부분의 경우 종속변량은 여러 개의 독립변량들과 관계를 갖고 있기 때문이다. 예를 들어, 단순선형회귀에서의 예제인 광고비와 판매액간의 관계에 있어서 판매액이 광고비에 의해 상당히 영향을 받지만 그 외에 제품의 품질등급, 판매매장의 개수와 크기 등에도 영향을 받을 것이라고 예상할 수 있다. 이와 같이 하나의 종속변량과 여러 개의 독립변량들 사이의 관계를 규명하고자 할 때 사용되는 통계적 방법이 **중선형회귀분석(multiple linear regression analysis)**이다. 그러나 단순선형회귀분석과 중선형회귀분석은 관련된 독립변량들의 개수만 다를 뿐이고 분석방법에는 별다른 차이가 없다.

중선형회귀모형

- 중선형 회귀모형에서는 종속변량 Y 와 k 개의 독립변량 X_1, \dots, X_k 가 다음과 같은 관계식을 가진다고 가정한다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \epsilon_i$$

즉, 종속변량은 각 독립변량의 일차함수로 나타내어지며 여기에 단순선형회귀모형에서와 같이 오차항을 나타내는 확률변량 ϵ 이 더해진다. 오차항 ϵ 에 대한 가정은 단순선형회귀에서의 가정과 같다. 위의 식에서 β_0 은 Y 축의 절편, β_i 는 Y 와 X_i 간의 기울기로써 다른 독립변량들이 고정되었을 때 X_i 가 Y 에 미치는 영향을 나타낸다.

- 일반적으로 중선형 회귀분석에서는 행렬과 벡터를 이용하면 식의 표현과 계산작업을 쉽게 할 수 있다. 이를테면, k 개의 독립변량이 있는 경우 관측점 $i=1,2,\dots,n$ 에서의 모집단 중선형회귀모형은 다음과 같은 간단한 식으로 나타내어진다.

$$Y = X\beta + \epsilon$$

여기에서

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

로 각각 정의된다.

회귀계수의 추정

- 중선형 회귀분석에서는 표본을 이용하여 $(k+1)$ 개의 회귀계수 $\beta_0, \beta_1, \dots, \beta_k$ 를 추정할 필요가 있다. 이 경우에도 오차들의 제곱의 합을 최소로 하는 최소제곱법을 사용한다. 즉, 오차제곱의 합

$$S = \sum_{i=1}^n \epsilon_i^2 = \epsilon' \epsilon = (Y - X\beta)'(Y - X\beta)$$

를 최소화하는 β 를 구한다. 단순선형회귀에서와 마찬가지로 위의 오차제곱합을 β 에 대해 미분하여 0으로 놓고 풀면 되는데, 이때 β 의 최소제곱추정량을 b 라 하면 b 는 다음의 정규방정식을 만족한다.

$$(X'X)b = X'Y$$

그러므로 만약 $X'X$ 의 역행렬이 존재하면 β 의 최소제곱추정량 b 는

$$b = (X'X)^{-1} X'Y$$

로 주어진다. (참고: 이 공식을 사용하면 계산오차가 크기 때문에 통계패키지 등

에서의 계산에 있어서는 다른 방법을 사용한다.)

- 표본을 이용하여 추정된 회귀계수를 $b=(b_0, b_1, \dots, b_k)$ 라 하면, 반응변량 Y 의 예측값은

$$\hat{Y}_i = b_0 + b_1 X_{i1} + \dots + b_k X_{ik}$$

로 주어지고, 잔차는

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_{i1} + \dots + b_k X_{ik})$$

로 주어지며, 역시 벡터를 사용하여 잔차벡터 e 를 다음과 같이 정의할 수 있다.

$$e = Y - Xb$$

회귀모형의 적합도와 분산분석

- 중선형 회귀분석에서도 추정된 회귀직선의 타당성을 조사하기 위해 잔차표준오차와 결정계수가 사용된다. 단순 선형회귀에서 이들 측도의 계산공식은 잔차, 즉, Y 의 관측값과 예측값의 함수로 주어지므로 독립변량의 개수와는 상관이 없었다. 그러므로, 중선형 회귀에서도 같은 공식을 사용할 수 있으며 단지 각 제곱합이 가지는 자유도의 값에 차이가 있다.
- 중선형 회귀분석에서 잔차표준오차는 다음과 같이 정의된다.

$$s = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

단순 선형회귀와의 차이는 잔차 $e_i = Y_i - \hat{Y}_i$ 를 계산하기 위해서는 $(k+1)$ 개의 회귀계수가 추정되어야 하므로 남아있는 자유도는 $(n-k-1)$ 이다. s^2 은 단순 선형회귀에서와 마찬가지로 잔차평균제곱(MSE)과 같은 통계량이다. 결정계수는 $R^2 = SSR/SSR$ 로 주어지며 그 해석은 단순 선형회귀에서와 같다.

- 제곱합도 단순 선형회귀에서와 같은 공식으로 정의되고, 대응되는 자유도와 함께 다음과 같이 분할될 수 있으며 이들을 이용한 분산분석표는 표 12.5에 주어진다.

제곱합: $SSR = SSE + SSR$
 자유도: $n-1 = (n-k-1) + k$

표 12.5 중선형회귀에서의 분산분석표

요인	제곱합	자유도	평균제곱	F값
처리 오차	SSR	k	$MSR = SSR / k$	$F_0 = MSR / MSE$
	SSE	$n-k-1$	$MSE = SSE / (n-k-1)$	
전체	SST	$n-1$		

- 위의 분산분석표에 주어진 F 값은 회귀식의 유의성 검정에 사용되는데 이 때의 귀무가설은 모든 독립변량은 종속변량과 선형관계가 없다는 것이다. 즉,

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: k\text{개의 } \beta_j \text{ 중 적어도 하나는 } 0 \text{ 이 아니다}$$

를 검정한다. 귀무가설 하에서 F_0 는 자유도 k 와 $(n-k-1)$ 인 F 분포를 따르므로 만약 $F_0 > F_{k, n-k-1; \alpha}$ 이면 H_0 를 유의수준 α 하에서 기각할 수 있다. 개개의 β 에 대한 검정도 할 수 있는데 이는 다음절에 설명되어 있다. (역시, 『eStat』에서는 이 검정에 대한 p -값을 계산하여 주므로 이 p -값을 이용하여 검정한다. 즉, p -값이 유의수준보다 작으면 귀무가설을 기각한다.)

중선형회귀분석에서의 추론

- 단순 선형회귀에서와 같이 중선형회귀에서도 관심의 대상이 되는 모수는 각 회귀계수 $\beta_0, \beta_1, \dots, \beta_k$ 와 Y 의 평균값이다. 이들 모수에 대한 추론은 점추정량의 확률분포를 구함으로써 가능하게 된다. 오차항 ε_i 들이 독립이고 모두 $N(0, \sigma^2)$ 의 분포를 가진다는 가정 하에서

$$b_i \sim N(\beta_i, c_{ii}\sigma^2), \quad i=0, 1, \dots, k$$

임을 보일 수 있다. 위에서 c_{ii} 는 $(k+1) \times (k+1)$ 행렬인 $(X'X)^{-1}$ 의 i 번째 대각원소이다. 그리고 모수 σ^2 대신에 추정량 s^2 을 사용하면 다음과 같이 t 분포를 이용하여 각 회귀계수에 대한 추론을 할 수 있다.

회귀계수 β_i 에 관한 추론

점추정량: b_i

점추정량의 표준오차: $SE(b_i) = \sqrt{c_{ii} \cdot s}$

신뢰구간: $\beta_i = b_i \pm t_{n-k-1; \alpha/2} \cdot SE(b_i)$

가설검정:

귀무가설: $H_0: \beta_i = \beta_0$

검정통계량: $t = \frac{b_i - \beta_0}{SE(b_i)}$

H_0 기각역: 대립가설이 $H_1: \beta_i < \beta_0$ 이면 $t < -t_{n-k-1; \alpha}$
 대립가설이 $H_1: \beta_i > \beta_0$ 이면 $t > t_{n-k-1; \alpha}$
 대립가설이 $H_1: \beta_i \neq \beta_0$ 이면 $|t| > t_{n-k-1; \alpha/2}$

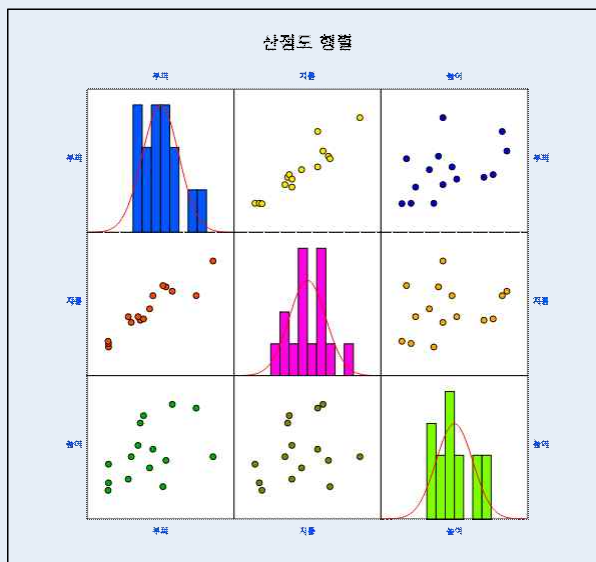
- 중선형 회귀분석에서의 잔차분석은 단순선형회귀에서와 동일하다.

[예 12.7] 산림지역에서 나무를 벌목할 때 해당 지역의 목재량을 조사할 필요가 있을 것이다. 그러나 나무의 부피를 직접 측정하는 것은 어렵기 때문에 상대적으로 측정이 쉬운 나무의 지름과 높이를 이용하여 부피를 추정하는 방법을 생각할 수 있다. 표 12.6의 자료는 어느 지역에서 15그루의 나무를 표본으로 추출하여 벌목한 후, 지름, 높이, 부피를 측정한 값들이다.(지름은 지상에서 1.5m 지점에서 측정되었다.) 이 데이터에 대한 산점도행렬을 그리고 이 문제에 대한 회귀모형을 생각해보자. (이 자료는 『eStat』의 Ex ⇒ 01Korean ⇒ 123중회귀_나무부피.csv 이름으로 저장되어 있다.)

표 12.6 나무의 지름, 높이, 부피 자료

지름(cm)	높이(m)	부피(m^3)
21.0	21.33	0.291
21.8	19.81	0.291
22.3	19.20	0.288
26.6	21.94	0.464
27.1	24.68	0.532
27.4	25.29	0.557
27.9	20.11	0.441
27.9	22.86	0.515
29.7	21.03	0.603
32.7	22.55	0.628
32.7	25.90	0.956
33.7	26.21	0.775
34.7	21.64	0.727
35.0	19.50	0.704
40.6	21.94	1.084

『eStat』으로 자료를 불러와 회귀분석 아이콘을 선택하여 나타나는 변량 선택박스에서 ‘Y 변량’을 부피 ‘by X변량’을 차례로 지름과 높이를 선택하면 [그림 12.16]과 같은 산점도행렬이 나타난다. 부피와 지름의 관련성이 높고, 부피와 높이, 지름과 높이도 어느 정도 관련성이 있음을 관찰할 수 있다.



[그림 12.16] 부피(Y)와 지름, 높이의 산점도행렬

기초통계량						
변량	변량명	자료수	평균	표준편차	표준오차	95% 신뢰구간
변량 1	부피	15	0.590	0.234	0.060	(0.461, 0.720)
변량 2	지름	15	29.407	5.514	1.424	(26.353, 32.460)
변량 3	높이	15	22.266	2.313	0.597	(20.985, 23.547)
결측수	0					

상관계수행렬				
상관계수 H ₀ : ρ=0 ρ≠0 t-값 p-값	변량명	변량 1	변량 2	변량 3
변량 1	부피	1	0.934 t-값 = 9.456 p-값 < 0.0001	0.464 t-값 = 1.889 p-값 0.0814
변량 2	지름	0.934 t-값 = 9.456 p-값 < 0.0001	1	0.263 t-값 = 0.984 p-값 0.3431
변량 3	높이	0.464 t-값 = 1.889 p-값 0.0814	0.263 t-값 = 0.984 p-값 0.3431	1

[그림 12.17] 부피(Y)와 지름, 높이의 기초통계량 및 상관계수행렬

나무의 지름과 높이를 이용하여 부피를 추정하고자 하는 것이므로, 부피가 종속변량 Y 가 되고, 지름과 높이가 각각 독립변량 X_1, X_2 가 되어 다음과 같은 회귀모형을 생각할 수 있다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \quad i=1,2,\dots,15$$

[그림 12.16]의 산점도행렬 밑의 선택사항에서 ‘상관 및 회귀분석’ 버튼을 누르면 결과저장창에 [그림 12.18]과 같은 추정된 회귀직선, 분산분석표 등을 구할 수 있다. 적합된 회귀식은 다음과 같다.

$$\hat{Y}_i = -1.024 + 0.037X_{1i} + 0.024X_{2i}$$

위에서 0.037은 지름 (X_1)이 1(cm)만큼 커질 때 증가하는 나무의 부피를 나타낸다. [그림 12.18]의 분산분석표에서 계산된 F 값 = 73.12에 대한 p -값이 0.0001보다 작으므로 유의수준 $\alpha=0.05$ 에서 귀무가설 $H_0: \beta_1 = \beta_2 = 0$ 을 기각할 수 있다. 그리고 결정계수는 $R^2=0.924$ 로 종속변량의 총변량 중 92.4% 정도가 회귀직선에 의해 설명되고 있다. 위의 두 가지 결과에 의해 우리는 나무의 지름과 높이가 부피의 추정에 상당히 유용하다고 결론 지을 수 있다.

회귀분석					
회귀선 $y =$	(-1.024) + (0.037) X_1 + (0.024) X_2				
중상관계수	0.961	결정계수	0.924	추정오차	0.069
모수	추정값	표준오차	t-값	p-값	95% 신뢰구간
β_0	-1.024	0.188	-5.458	0.0001	(-1.358, -0.689)
β_1 지름	0.037	0.003	10.590	< 0.0001	(0.031, 0.043)
β_2 높이	0.024	0.008	2.844	0.0148	(0.009, 0.038)
[분산분석]					
요인	제곱합	자유도	평균제곱	F-값	p-값
회귀	0.7058	2	0.3529	73.1191	< 0.0001
오차	0.0579	12	0.0048		
전체	0.7638	14			

[그림 12.18] 나무자료에 대한 중선형 회귀분석의 분산분석표

결과화면 [그림 12.18]에서 $SE(b_1)=0.003$, $SE(b_2)=0.008$ 로 주어지고, $t_{12;0.025}=2.179$ 이므로 각 회귀계수에 대한 신뢰구간은 다음과 같이 주어진다. 결과 화면과 수치가 다른 것은 소숫점 이하 계산의 오차이다.

$$\beta_1 \text{의 } 95\% \text{ 신뢰구간: } 0.037 \pm (2.179)(0.003)$$

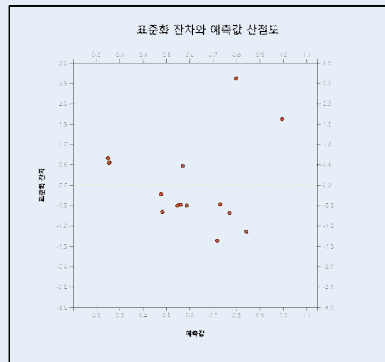
$$\Rightarrow (0.029, 0.045)$$

$$\beta_2 \text{의 } 95\% \text{ 신뢰구간: } 0.024 \pm (2.179)(0.008)$$

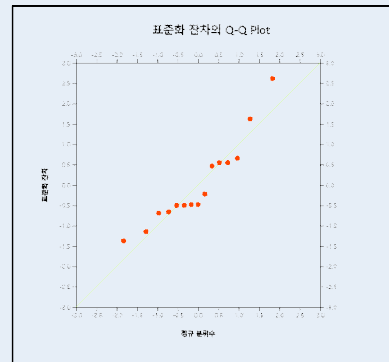
$$\Rightarrow (0.006, 0.042)$$

그리고 가설 $H_0: \beta_i=0$, $H_1: \beta_i \neq 0$, $i=1,2$ 의 검정에서 각 p -값은 유의수준 0.05보다 작으므로 각각의 귀무가설을 기각할 수 있다.

표준화 잔차의 산점도는 [그림 12.19]와 같고 Q-Q 산점도는 [그림 12.20]과 같다. 표준화 잔차의 산점도에서 특별한 경향은 나타나지 않으나 1개 정도의 이상값이 나타나고 있으며, Q-Q 산점도는 정규성 가정이 어느 정도 잘 만족됨을 보여준다고 할 수 있다.



[그림 12.19] 나무자료에 대한 중선형 회귀분석의 잔차분석



[그림 12.20] 나무자료에 대한 중선형 회귀분석의 Q-Q 산점도