제 6 장 회귀분석

6.1 서론 및 용어
6.2 LSE
6.3 회귀모형의 MLE
6.4 SLR의 분석
6.5 MLR의 분석
6.6 상관관계 분석

§6.1 서론 및 용어

회귀(regression)라는 용어의 유래는 다음과 같다. 유전학자 Galton(1822-1911)의 연구 결과에 의하면, 키가 작은 (큰) 아버지를 가진 아들의 키는 평균치보다 작지만 (크지만) 아버지의 키보다는 커서 (작아서) 평균치 쪽으로 "회귀"하는 경향이 있다고 한다(문헌 [4] 참조). 오늘날에는 2개 이상의 변수들간의 관계식을 찾아내고, (이는 추정에 해당됨) 이 관계식의 타당성과 정확성을 검토하는 (이는 검정에 해당됨) 통계적 방법을 회귀분석(regression analysis)이라 한다.

변수들 간의 관계식이 직선 또는 평면인 모형을 선형(linear) 회귀모형이라 하고 곡선 또는 곡면인 모형을 비선형(nonlinear) 회귀모형이라 하는데, 이 책에서는 선형 회귀모형만 다룬다.(§6.5.5 참조). 그리고, 직선 관계식을 회귀직선(regression line)이라 하고, 평면 관계식을 회귀평면(regression plane)이라 한다. 또한, 회귀직선에 관련된 추정 및 검정은 SLR(simple linear regression)이라 하고, 회귀평면에 관련된 추정 및 검정을 MLR(multiple linear regression)이라 한다.

아래의 표는 아버지와 아들의 키를 (㎝ 단위로) 나타낸 것이다.

i	1	2	3	4	5	6	7
아버지의 키 (x_i)	156	159	168	177	183	183	192
아들의 키 (y_i)	166	166	169	175	183	186	187

$(x_i, y_i), i=1,\cdots,7$은 xy 평면상의 7개의 점에 해당된다. 이 경우, 회귀직선은 $y=60+0.6x$ 인데, (§6.2.1 참조), 이 직선은 주어진 7개의 점에 가장 잘 들어맞는 직선이다.

<비고 6.1.1> "가장 잘 들어맞는 직선"이란 "주어진 점에서 직선까지의 수직 방향 (즉, y방향) 거리를 제곱한 값들의 총합을 최소가 되게 하는 직선"을 의미한다.

<비고 6.1.2> x_i의 평균은 $\overline{x}=174$ 이고 y_i의 평균은 $\overline{y}=176$ 인데, 회귀직선 $y=60+0.6x$는 점 $(\overline{x},\overline{y})$를 통과한다.

<비고 6.1.3> Galton의 연구결과는 회귀직선의 기울기가 0보다 크고 1보다 작음을 의미한다.

§5.6에서 SLR 모형을 식 (5.6.4)로 표현했다. 즉,

$$Y_x=\beta_0+\beta_1 x+\varepsilon_x,\ \varepsilon_x\sim iid\ N(0,\sigma^2) \tag{6.1.1}$$

인데, 회귀직선의 계수인 60과 0.6은 각각 β_0와 β_1에 대한 최우추정치이다.

<비고 6.1.4> 회귀직선의 계수는 통계학적으로 최우추정치이지만 기하학적으로는 <비고 6.1.1>에 의한 것이다. 이에 따라, 회귀직선의 계수를 MLE인 동시에 LSE(least squares estimate)라 부른다. 또한, LSE의 "E"는 estimate 뿐만 아니라 estimator와 estimation을 의미하기도 한다 (<비고 1.6.1> 참조).

그리고, (§3.5.1에 등장한) MLE의 불변성(invariance)에 의해서 $(60+0.6x)$는 $(\beta_0+\beta_1 x)$에 대한 최우추정치인데, $(\beta_0+\beta_1 x)$는 바로 $E(Y_x)$이다. (비고: 식 (6.1.1)에서 $E(\varepsilon_x)=0$ 이므로 $E(Y_x)=\beta_0+\beta_1 x$임.)

회귀직선 $y=60+0.6x$가 $E(Y_x)$에 대한 최우추정치라면 Y_x는 과연 무엇인가? 식 (6.1.1)에 의하면, 키가 x인 아버지를 가진 아들의 키가 Y_x이다. 그러나, 키가 x인 아버지가 여럿인 경우에 (또는, 아버지가 같더라도), 아들들의 키가 모두 같지는 않다. 즉, Y_x는 상수가 아니라 확률변수인데, 식 (6.1.1)은 바로 Y_x가 평균이 $(\beta_0+\beta_1 x)$이고 분산이 σ^2인 정규분포를 따른다는 가정인 셈이다. 그렇다면, 무엇이 모집단이고 무엇이 표본인가? 사실 Y_x의 분포가 바로 모분포이다. 그러나, x값이 다르면 (모평균 "$\beta_0+\beta_1 x$"가 달라지기 때문에) 모분포도 다르고 또한 모집단도 다르다. 예제에서는 모두 6개의 모집단이 등장하고 (비고 : $x_5=x_6=188$), 관찰된 표본의 수 역시 6개인데 (<비고 5.2.1> 참조), 이 중에서 크기가 1인 표본은 5개이고 크기가 2인 표본은 하나이다.

회귀분석의 첫 단계는 회귀직선 또는 회귀평면을 추정하는 것이다. 그리고, 둘째 단계

에서는 $\beta_0, \beta_1, \cdots$에 대한 검정을 한다. 또는 $\beta_0, \beta_1, \cdots$을 하나로 묶어서 회귀모형 자체에 대한 타당성(validity)을 검정하기도 한다. 그러나, 첫째와 둘째 단계는 준비작업일 뿐이고, 회귀분석의 주목적은 예측(prediction)이다. 예를 들어 아버지의 키가 180인 경우 아들의 키에 대한 예측치는 $60+(0.6)(180)=168$이다. 또한, 아버지의 키가 174이면 아들의 키의 예측치는 176이다.(<비고 6.1.2) 참조). 물론, 예측의 정확성을 검정하는 것도 회귀분석에 포함된다.

<비고 6.1.5> 회귀분석에서의 예측은 “prediction"이고, 시계열분석(time-series analysis)에서의 예측은 ”forecasting" 임.

x와 Y_x에 대한 호칭은 여러 가지가 있다. (비고: 식 (5.6.5)의 MLR 모형에서는 x가 벡터 $(x_1, x_2, \cdots)$를 의미함.) 첫째, 아들의 키가 크고 작음이 아버지의 키에 의해서 설명된다는 의미에서 x를 설명변수(explanatory variable)라 하고, x에 반응하는 것이 Y_x라는 의미에서 Y_x를 반응변수(response variable)라 한다. 둘째로, x를 회귀변수 또는 예측변수라 하고 Y_x를 피회귀변수 또는 피예측변수라 하기도 한다. 그러나, 가장 흔히 사용되는 호칭은 독립변수(x)와 종속변수(Y_x)이다. 이는 Y_x가 x의 함수라는 점을 강조하는 것이기도 하지만, 이때 “독립”이라는 표현은 x가 확률변수가 아니라는 점을 암시하는 것이기도 하다. x는 (확률변수가 아닐뿐더러) 많은 경우에 제어변수(control variable)의 역할까지 한다. 예를 들어, x는 광고비이고 Y_x는 매출액이라 하자. 매출액을 직접 결정할 수는 없다. 다만, 직접 결정할 수 있는 광고비를 통해서 (즉, 광고비를 제어 또는 조절함으로써) 간접적으로 매출액에 영향을 끼칠 수 있다.

독립변수가 하나일 때, $E(Y_x)=\beta_0+\beta_1 x$에 대한 추정치(또는 추정식)를 회귀직선이라 했다. 그리고, 독립변수가 둘이면 $E(Y_x)=\beta_0+\beta_1 x_1+\beta_2 x_2$에 대한 추정치는 회귀평면이 된다. 그런데, 독립변수가 셋 이상인 경우에도 여전히 $E(Y_x)$에 대한 추정치를 회귀평면이라 한다 (엄격히 하자면 평면이 아니라 초평면(hyperplane)임).

§6.2 LSE

6.2.1 SLR에 대한 LSE

§6.1에서 주어진 7개의 점에 가장 잘 들어맞는 직선은 $y=60+0.6x$라고 했는데 이

를 먼저 확인해 보자 (<비고 6.1.1> 참조).

회귀직선을 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$라 하면

$$f_i \equiv y_i-\hat{y}_i \equiv y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i) \tag{6.2.1}$$

는 점 (x_i, y_i)로부터 회귀직선까지의 수직 방향 (또는 y방향) 거리를 나타낸다. 그리고, 이들의 제곱합(SS: sum of squares)을 SSE라 하자(§6.3.1 마지막 문단 참조). 즉,

$$SSE=\sum_i f_i^2=\sum_i \{ y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)\}^2 \tag{6.2.2}$$

이다. 그러면, LSE의 정의에 의해 (<비고 6.1.4> 참조), $\hat{\beta}_0$과 $\hat{\beta}_1$은 아래의 식을 만족시킨다.

$$\frac{\partial SSE}{\partial \hat{\beta}_0}=-2\sum_i f_i=0 \tag{6.2.3}$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1}=-2\sum_i x_i f_i=0 \tag{6.2.4}$$

식 (6.2.3)으로부터는 다음의 관계식을 쉽게 얻을 수 있다 (<비고 6.1.2> 참조).

$$\bar{y}=\hat{\beta}_0+\hat{\beta}_1 \bar{x} \tag{6.2.5}$$

반면에, 식 (6.2.4)는 약간의 손질이 필요한데 결과는 다음과 같다.

$$\hat{\beta}_1=\frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} \tag{6.2.6}$$

§6.1의 예제에서는 $\bar{x}=174$, $\bar{y}=176$, $\sum_{i=1}^{7}(x_i-\bar{x})(y_i-\bar{y})=720$, $\sum_{i=1}^{7}(x_i-\bar{x})^2=1080$ 이므로, 먼저 $\hat{\beta}_1=720/1080=2/3$을 얻고 나서 이를 식 (6.2.5)에 대입하면 $\hat{\beta}_0=60$을 얻는다.

<비고 6.2.1> 수직방향 거리는 최단거리가 아니다. 최단거리의 제곱합을 최소가 되게하는

직선은 직교(orthogonal)회귀직선이라 하는데, 이는 다변량(multivariate) 분석의 범주에 속한다(§6.6.4 참조).

6.2.2 LSE의 역학적 해석

LSE 방법으로 얻은 회귀직선은 힘의 평형으로 해석할 수 있다(문헌[3]참조). 회귀직선을 단단한 막대기라 하고, 식 (6.2.1)의 f_i를 막대기에 작용하는 힘이라 하자. 즉, 막대기의 $(x_i, \hat{y}_i)$ 지점에 크기가 f_i인 힘이 수직방향으로 작용한다고 하자. (비고: $f_i>0$ 이면 막대기를 위로 잡아 당기고, $f_i<0$ 이면 아래로 잡아 당김.)

막대기가 (움직이지 않고) 평형상태에 있을 조건은 두 가지이다. 첫째는 $\sum_i f_i = 0$ 인데, 이는 식 (6.2.3)에 해당된다. 합력이 0이면 최소한 막대기의 중심인(?) $(\bar{x}, \bar{y})$는 고정된다 (<비고 6.1.2> 참조). 그러나, 여전히 $(\bar{x}, \bar{y})$를 축으로 회전운동은 할 수 있다. 이러한 회전운동을 방지하기 위한 조건은 바로 식 (6.2.4)인데, 이를

$$\sum_i (x_i - \bar{x}) f_i = 0 \tag{6.2.7}$$

으로 고치면 이해하기 쉽다. (비고: 식 (6.2.7)은 식 (6.2.3)과 (6.2.4)로부터 얻음.) 식 (6.2.7)에서 $(x_i - \bar{x}) f_i > 0$ 이면 시계반대방향으로 그리고 $(x_i - \bar{x}) f_i < 0$ 이면 시계방향으로 회전효과(torque)가 작용하는데, 이들의 합이 0이면 (회전하지 않고) 평형상태가 된다.

6.2.3 MLR에 대한 LSE

먼저, 독립변수가 두 개인 경우를 다룬다. 주어진 점 $(x_{i1}, x_{i2}, y_i), i = 1, \cdots, n$는 이제 3-차원 공간에 있는 n개의 점이 되고, 회귀평면 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$는 3-차원 공간에서 2-차원을 차지한다.

점 (x_{i1}, x_{i2}, y_i)로부터 회귀평면까지의 수직방향 (또는 y방향) 거리는

$$f_i \equiv y_i - \hat{y}_i \equiv y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) \tag{6.2.8}$$

가 된다. 그리고, LSE인 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$는 여전히 $SSE \equiv \sum_i f_i^2$을 최소가 되게 한다. LSE의 조건식은 이제

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_0}=\sum_i f_i=0 \tag{6.2.9}$$

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_1}=\sum_i f_i x_{i1}=0 \tag{6.2.10}$$

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_2}=\sum_i f_i x_{i2}=0 \tag{6.2.11}$$

인데, 이들에 대한 역학적 해석은 다음과 같다. 회귀평면을 딱딱한 널빤지라 하면, 식 (6.2.9)는 널빤지에 작용하는 (수직방향) 힘의 합이 0임을 의미한다. 또한 기하학적으로는 점 $(\overline{x_1}, \overline{x_2}, \overline{y})$가 널빤지 상에 있음을 의미한다 (<비고 6.1.2> 참조). 합력이 0이더라도 회전운동은 가능한데, 이를 방지하기 위한 조건이 식 (6.2.10)과 (6.2.11)이다. 구체적으로 식 (6.2.10)은 널빤지가 x_2축(axis)을 축(pivot)으로 회전하는 것을 방지하고, 식 (6.2.11)은 x_1축(axis)을 축(pivot)으로 회전하는 것을 방지한다.

일반적으로, 독립변수가 k개인 경우에

$$f_i \equiv y_i - \hat{y}_i \equiv y_i - (\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j),\ i = 1, \cdots n, \tag{6.2.12}$$

이라 하면, LSE인 $\hat{\beta}_0$와 $\hat{\beta}_j,\ j = 1, \cdots, k$를 구하는 방정식은

$$\begin{aligned} &\sum_{i=1}^{n} f_i = 0 \\ &\sum_{i=1}^{n} f_i x_{ij} = 0,\ \ j=1,\cdots,k \end{aligned} \tag{6.2.13}$$

인데, 이를 벡터와 행렬로 표현하면 풀기가 쉽다.

다음과 같이 모든 벡터는 열(column)벡터로 정의한다.

$$f \equiv \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix},\ \ y \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \ \hat{y} \equiv \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix},\ \ \hat{\beta} \equiv \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \tag{6.2.14}$$

그리고, 행렬 X는 다음과 같이 정의한다.

$$X \equiv \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{6.2.15}$$

그러면, 식 (6.2.12)과 (6.2.13)은 각각

$$f = y - \hat{y} = y - X\hat{\beta} \tag{6.2.16}$$

$$X'f = 0 \tag{6.2.17}$$

이 된다. (비고: 식 (6.2.17)에서 X'은 X의 transpose이고 우변의 0은 0벡터를 의미함.) 따라서 식 (6.2,16)을 식 (6.2.17)에 대입하면

$$X'y - X'X\hat{\beta} = 0 \tag{6.2.18}$$

인데, 이를 $\hat{\beta}$에 대해서 풀면 다음을 얻는다.

$$\hat{\beta} = (X'X)^{-1}(X'y) \tag{6.2.19}$$

참고로 $SSE = \sum_{i=1}^{n} f_i^2 = f'f$에 식 (6.2.19)를 대입(하여, 간단히)하면 다음 식을 얻는다.

$$SSE = y'y - \hat{\beta}'X'y \tag{6.2.20}$$

6.2.4 MLR 예제

문헌 [9]의 예제 11.12는 다음과 같다.

$$y = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \tag{6.2.21}$$

먼저, 식 (6.2.6)의 분자와 분모에 해당되는 $X'y$와 $(X'X)^{-1}$는

$$X'y=\begin{bmatrix}5\\7\\13\end{bmatrix},\quad (X'X)^{-1}=\begin{bmatrix}17/35 & 0 & -1/7\\ 0 & 1/10 & 0\\ -1/7 & 0 & 1/14\end{bmatrix} \tag{6.2.22}$$

이다. 다음, 이들을 식 (6.2.19)에 대입하면

$$\hat{\beta}=\begin{bmatrix}4/7\\7/10\\3/14\end{bmatrix}\approx\begin{bmatrix}0.5714\\0.7000\\0.2143\end{bmatrix} \tag{6.2.23}$$

이므로, 회귀평면은

$$\hat{y}\approx 0.571+0.7\,x_1+0.214\,x_2 \tag{6.2.24}$$

가 된다. 또한, 식 (6.2.20)으로부터 다음을 얻는다.

$$SSE\approx 0.4571 \tag{6.2.25}$$

§6.3 회귀모형의 MLE

6.3.1 LSE와 MLE

§6.2에서 구한 LSE가 MLE와 동일함을 보인다(<비고 6.1.4>참조). §6.2에서와 같이 x_{ij}를 j번째 ($j=1,\cdots,k$) 독립변수의 i번째 ($i=1,\cdots,n$) 관찰치라 하고, y_i를 종속변수의 i번째 관찰치라 하자. 이에 따라, §6.1에서 Y_x로 표현하던 종속변수를 지금부터는 Y_i로 표기한다. 그러면, 회귀모형은

$$Y_i = \mu_i + \varepsilon_i,\ i=1,\cdots,n \tag{6.3.1}$$

$$where\ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{6.3.2}$$

인데, 확률변수 ε_i의 분포에 대한 가정은 $iid\ N(0,\sigma^2)$이므로 결국

$$Y_i \sim N(\mu_i, \sigma^2),\ i=1,\cdots,n \tag{6.3.3}$$

이 된다.

<비고 6.3.1> $Y_1,\cdots,Y_n$은 (평균이 다르므로) 동일하지는 않지만, ($\varepsilon_1,\cdots,\varepsilon_n$이 iid 확률변수이므로) 서로 독립이다.

$Y_1,\cdots,Y_n$이 독립이므로, 이들의 결합밀도함수인 LF는

$$L = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2} \tag{6.3.4}$$

이다 (식 (3.2.1) 참조). 그리고, 식 (6.3.4)에 자연대수를 취하면

$$\ln L = K - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 \tag{6.3.5}$$

가 된다 (식 (3.2.2)이하 부분 참조). 이때, 유의할 점은 다음과 같다. 식 (6.3.5)를 $\beta_0,\beta_1,\cdots,\beta_k$에 대해서 (비고: $\mu_1,\cdots,\mu_n$은 $\beta_0,\beta_1,\cdots,\beta_k$의 함수임. 식 (6.3.2) 참조) 편미분한 식을 (0으로 놓고) 푸는 대신에, 마지막 항에 있는

$$\sum_{i=1}^{n} (y_i - \mu_i)^2 \tag{6.3.6}$$

를 편미분한 식을 풀어도 같은 결과를 얻는다. 즉, 식 (6.3.5)를 최대가 되게하는 $\beta_0, \beta_1, \cdots, \beta_k$ 값들을 식 (6.3.6)을 최소가 되게하는 $\beta_0, \beta_1, \cdots, \beta_k$와 동일하다. 그런데, LSE는 바로 식 (6.3.6)이 최소가 되는 조건식을 풀어서 얻은 것이다 (식 (6.2.8) 참조). 따라서, $\beta_0, \beta_1, \cdots, \beta_k$에 대한 MLE를 $\widehat{\beta}_0, \cdots, \widehat{\beta}_k$라 하면 이는 MLE인 동시에 LSE이다. (비고: MLE의 표기인 $\widehat{\beta}_0, \cdots, \widehat{\beta}_k$를 편의상 LSE의 표기로도 사용했음.)

<비고 6.3.2> 모분포에 대한 가정이 없는 LSE 방법은 일종의 heuristic 방법이다 (<비고 3.6.1> 참조). 따라서, 정규분포의 가정 하에 얻은 MLE가 LSE와 일치한다는 사실은 MLE의 robustness를 뒷받침하는 것이라 할 수 있다 (<비고 3.6.1>의 윗 문단 참조).

6.3.2 *SSE*

MLE인 $\widehat{\beta}_0, \cdots, \widehat{\beta}_k$를 식 (6.3.6)에 대입하면 §6.2에서 *SSE*라 부르던 것이 된다. 즉,

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 \tag{6.3.7}$$

$$where\ \widehat{\mu}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik} \tag{6.3.8}$$

이다. 그리고, 이를 식 (6.3.5)에 대입한 다음 σ^2에 대해서 편미분하면, σ^2에 대한 MLE로

$$\widehat{\sigma^2} = \frac{SSE}{n} \tag{6.3.9}$$

를 얻는다 (식(3.2.4)참조).

이제 용어에 대해서 한가지 짚고 넘어갈 때가 되었다. 회귀분석의 주목적은 예측이라 했다(<비고 6.1.5> 참조). 독립변수의 값이 x_{ij}일 때 ($i = 1, \cdots, n$, $j = 1, \cdots, k$), Y_i에 대한 예측치로 식 (6.3.8)의 $\widehat{\mu}_i$를 사용한다. (비고: $\widehat{\mu}_i$는 식 (6.2.12)의 $\widehat{y}_i$와 같음.) 이때, 실제 관측치 y_i와 예측치 $\widehat{\mu}_i$의 차이인 $(y_i - \widehat{\mu}_i)$를 잔차(residual)라 부른다. (비고:

§6.2에서는 잔차를 f_i로 포기하고, 이를 힘으로 해석했음.) 이에 따라, 식 (6.3.7)을 잔차제곱합(residual SS)이라 부르는 책이 많다. 그런데도 이 책에서 식 (6.3.7)을 SSE라 부르는 이유는 회귀모형의 SSE가 ANOVA의 SSE와 동일한 역할을 하기 때문이다.

사실 SSE는 모든 선형모형에서 동일한 역할을 한다. 즉, 모든 선형모형에서 σ^2에 대한 MLE는 SSE/n 이다. 예를 들어, 식 (5.6.3)에서 σ^2에 대한 MLE는 식 (4.4.8)인데, 이때 SSE는 바로 3장에서 SS라 불렀던 것이다 (<비고 3.4.1> 참조).

이제, SSE의 의미와 용도를 더욱 확장시킨다. 지금까지 SSE라 부른 것은 5장의 ANOVA에서 정의된 SSE이다. 즉, 전체 Variation인 TSS 중에서 "선형모형"에 의해서 설명된 부분을 빼고 남은 부분을 SSE라 불렀는데, 이때 "선형모형"이라 함은 CM을 의미한다 (<비고 5.6.1> 참조). 예를 들어, TWA의 SSE는 식 (5.6.1) 하에서 "설명안된 부분"인데 이를 SSE_{CM}이라 하자. 반면에, TWA의 RM인 식 (5.6.2) 하에서 "설명안된 부분"을 SSE_{RM}이라 하면

$$SSE_{RM} = SSE_{CM} + SST_r \tag{6.3.10}$$

의 관계가 성립한다. 즉, 식 (5.5.8)은 RM 하에서 σ^2에 대한 MLE인 반면에, 식 (5.5.9)는 CM 하에서 σ^2에 대한 MLE이다. 따라서, 식 (5.5.10)은

$$\frac{(SSE_{RM} - SSE_{CM})/(d_{RM} - d_{CM})}{SSE_{CM}/d_{CM}} \sim F(d_{RM} - d_{CM}, d_{CM}) \tag{6.3.11}$$

으로 표현할 수 있다. (비고: <비고 5.6.4>의 경우에는 $SSE_{RM} = SSE_{CM} + SSB$.)

6.3.3 회귀모형과 LRT

회귀모형을 포함한 모든 선형모형에서

$$\begin{aligned} H_0 &: RM \\ H_a (\cup H_0) &: CM \end{aligned} \tag{6.3.12}$$

에 대한 검정통계량은 식 (6.3.11)이다. TWA를 예로 들어서 설명했던 식 (6.3.11)을 이제 회귀모형으로 설명한다. 회귀모형의 LF는 식 (6.3.4)인데, LR은 여전히 식 (5.2.7)의 형태인

$$\lambda=(\frac{SSE_{CM}/n}{SSE_{RM}/n})\leq k \tag{6.3.13}$$

이다. 즉, 식 (5.2.7)에서 $\widehat{\sigma}_2$은 CM 하에서 σ^2에 대한 MLE이고, $\widehat{\sigma}_0^2$는 RM 하에서 σ^2에 대한 MLE이다.

회귀모형에서 CM은 식 (6.3.1)이다. 반면에, RM은 다양하게 정의할 수 있다. 한마디로, $\{\beta_1,\cdots,\beta_k\}$ 중에서 하나 이상을 0으로 놓으면 RM이 된다. 구체적으로, $\{\beta_1,\cdots,\beta_k\}$ 중에서 하나만 0으로 놓으면 식 (6.3.13)은 T-test가 되고, 둘 이상을 0으로 놓으면 F-test가 된다.

<비고 6.3.3> 일반적으로, $\beta_1,\cdots,\beta_k$의 (함수를 0으로 놓는) 제약식의 개수가 1이면 T-test 이고 2 이상이면 F-test 인데, 이 책에는 "$\beta_j=0$" 형태의 제약식만 등장함.

<비고 6.3.4> 제약식이 $\beta_1=\cdots=\beta_k=0$ 인 경우의 RM을 "Null Model"이라 하고, 이에 대한 F-test를 "F-test for Model"이라 한다.

Null Model 하에서 SLR의 회귀직선은 수평선이 되고, MLR의 회귀평면은 수평면이 된다. 그리고, 회귀직선 또는 회귀평면은 여전히 점 $(\overline{x_1},\cdots,\overline{x_k},\overline{y})$를 포함한다(<비고 6.1.2> 참조). 따라서, Null Model 하에서의 SSE를 SSE_0라 하면

$$SSE_0\equiv\sum_{i=1}^{n}(y_i-\overline{y})^2 \tag{6.3.14}$$

인데, 이는 종전에 TSS라 부르던 것이다. 즉, Null Model 하에서 σ^2에 대한 MLE는 SSE_0/n이고, σ^2에 대한 MVUE는 $SSE_0/(n-1)$이다. 또한, $SSE_0/\sigma^2\sim\chi^2(n-1)$ 이므로, RM이 Null Model인 경우 식 (6.3.11)의 d_{RM}은 $(n-1)$이다.

<비고 6.3.5> 제약식이 모두 "$\beta_j=0$" 형태인 경우 (<비고 6.3.3> 참조) CM과 RM 하에서의 독립변수의 개수를 각각 k와 k'이라 하면, 식 (6.3.11)에서 "$d_{CM}=n-(k+1)$"이고 "$d_{RM}=n-(k'+1)$"이다. 따라서 $(d_{RM}-d_{CM})$

$=(k-k')$ 인데, $(k-k')$은 바로 RM 하에서의 제약식의 개수이다.

앞으로 자주 사용될 MSE는 바로 식 (6.3.11)의 분모를 의미한다. 즉,

$$MSE \equiv SSE_{CM}/(n-k-1) \tag{6.3.15}$$

이다.

6.3.4 $\hat{\beta}_i$의 분포

식 (6.3.11) 하나로 회귀분석에 관련된 모든 검정을 할 수 있다. 그러나, 여전히 $\hat{\beta} \equiv (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$의 확률분포가 필요하다. (비고: 행(row)벡터의 transpose는 열(column)벡터임.)

식 (6.2.19)의 $\hat{\beta}$은 $\beta=(\beta_0, \beta_1, \cdots, \beta_k)'$에 대한 점추정치이므로, $y=(y_1, \cdots, y_n)'$을 $Y=(Y_1, \cdots, Y_n)'$으로 대체하면 점추정량이 된다. 점추정량의 분포는 첫째로 β_j에 대한 신뢰구간을 구할 때 필요하다. 둘째로, 소위 예측구간(prediction interval)을 구하기 위해서는 $\hat{\beta}_0, \cdots, \hat{\beta}_k$간의 공분산까지 필요하다. 셋째로, 신뢰구간 및 예측구간 같은 구간추정을 할 때뿐만 아니라, 검정을 할 때에도 $\hat{\beta}$의 분포를 알면 식 (6.3.11)의 검정통계량을 간단히 얻을 수 있다.

β에 대한 점추정량인 $\hat{\beta}=(X'X)^{-1}(X'Y)$는 한마디로 $Y_1, \cdots, Y_n$의 (선형)함수이다. 따라서, <비고 6.3.1>과 <비고 2.15.1>에 의해서

$$\hat{\beta} \sim MVN(\beta, (X'X)^{-1}\sigma^2) \tag{6.3.16}$$

임을 보일 수 있다(증명은 생략함). 식 (6.3.16)에서 "MVN"은 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$의 결합분포가 "Multivariate Normal" 분포임을 의미한다. 따라서, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$은 각각 정규분포를 따른다(§2.2.3 참조). 그리고, $E(\hat{\beta})=\beta$이므로, $\hat{\beta}_0, \cdots, \hat{\beta}_k$는 모두 불편추정량이다. 반면에, $(X'X)^{-1}\sigma_2$은 공분산행렬(covariance matrix)이라는 것인데, 대각선(diagonal) 요소는 차례대로 $V(\hat{\beta}_0), V(\hat{\beta}_1), \cdots, V(\hat{\beta}_k)$이고 나머지는 $\hat{\beta}_0, \cdots, \hat{\beta}_k$간의 공분산이다.

또한, $\beta_0, \cdots, \beta_k$의 선형함수를

$$\sum_{j=0}^{k} a_j \beta_j = (\ a_0\ \ a_1 \cdots\ a_k\ ) \begin{vmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{vmatrix} \equiv a'\beta \tag{6.3.17}$$

로 표현하면, MLE의 불변성(invariance)에 의해서 $a'\beta$에 대한 MLE는 $a'\hat{\beta}$인데 이의 분포는

$$a'\hat{\beta} \sim N(a'\beta,\ a'(X'X)^{-1}a\sigma^2) \tag{6.3.18}$$

임을 보일 수 있다(증명은 생략함). (비고: $\hat{\beta}$가 $Y_1, \cdots, Y_n$의 선형함수이므로, $a'\hat{\beta}$ 역시 $Y_1, \cdots, Y_n$의 선형함수임.)

<비고 6.3.6> LSE인 동시에 MLE인 $\hat{\beta}$은 MVUE이기도 하다. 그런데, $\hat{\beta}$이 $Y_1, \cdots, Y_n$의 선형함수임을 강조하기 위해서, $\hat{\beta}$을 BLUE(best linear unbiased estomator)라 부르기도 한다. 또한, <비고 2.15.2>에서 $\overline{Y}$와 $\sum_{i=1}^{n} (Y_i - \overline{Y})^2$이 서로 독립이듯이, $\hat{\beta}$와 SSE_{CM}은 서로 독립이다. 그리고, 식 (6.3.15)의 MSE는 σ^2에 대한 MVUE이다.

§6.4 SLR의 분석

6.4.1 SLR과 LRT

§6.1의 예제는 SLR에서 $n=7$ 인 경우인데, §6.2.1에서 (β_0, β_1)에 대한 추정치로 $(60, 0.5)$을 얻었다. 따라서, $(\hat{\beta}_0, \hat{\beta}_1) = (60, 0.5)$을 식 (6.2.2)에 대입하면 $SSE_{CM}=40$ 을 얻고, 이를 식 (6.3.15)에 대입하면 $MSE=40/5=8$ 을 얻는다.

SLR에서 CM은 다음과 같다 (식 (6.3.1) 참조).

$$Y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i,\ i=1,\cdots,n \tag{6.4.1}$$

SLR에서 유의할 점은 Null Model인 (<비고 6.3.4> 참조)

$$Y_i = \beta_0 + \varepsilon_i,\ i=1,\cdots,n \tag{6.4.2}$$

가 유일한 RM이라는 점이다. 따라서, 식 (6.3.14)로부터 $SSE_0 = \sum_{i=1}^{7}(y_i - 176) = 520$ 을 얻는다. 그리고, 이들을 식 (6.3.11)에 대입하면

$$f = \frac{(520-40)/1}{40/5} = 60 \tag{6.4.3}$$

을 얻는다. 그런데, 분자 자유도가 1이고 분모 자유도가 5인 F-test에서, $\alpha=5\%$ 와 $\alpha=0.5\%$ 에 대한 UTT 기각역은 각각 6.61과 22.78이므로, 귀무가설인 RM을 (식 (6.3.12)참조) $\alpha=0.5\%$ 에서조차 기각할 수 있다.

통계 패키지에 의한 ANOVA Table은 다음과 같다.

Source of Variation	SS	자유도	MS	F	$PR>F$	R-SQUARE
Model	480	1	480	60	0.0006	0.9231
Error	40	5	8			
Total	520	6				

위의 Table에서 "$PR>F$"는 p-value이다(§4.6.3 참조). 그리고, "R-SQUARE"는

$$R^2 \equiv \frac{SSM}{TSS} = \frac{480}{520} = 0.9231 \tag{6.4.4}$$

인데, 이는 전체 Variation인 TSS(또는 SSE_0) 중에서 회귀모형에 의해서 설명된 부분인 SSM(SS for Model)이 차지하는 비율이다. (비고: $SSM = SSE_0 - SSE_{CM}$.) 따라서, 전체 Variation 중에서 92.31%가 회귀모형에 의해서 설명되었다고 할 수 있다.

6.4.2 β_1에 대한 추론

SLR 경우 식 (6.3.16)의 $(X'X)^{-1}$는 다음과 같다(계산은 생략함).

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{bmatrix} \tag{6.4.5}$$

따라서, $\hat{\beta}_1$의 분포는

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2) \tag{6.4.6}$$

인데, σ^2을 식 (6.3.15)의 MSE로 대체하면 (<비고 6.3.6> 참조)

$$T_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathrm{MSE} / \sum (x_i - \bar{x})^2}} \tag{6.4.7}$$

를 얻는다. (비고: t분포의 자유도 $(n-2)$는 MSE의 자유도임.)

먼저, 식 (6.4.7)을 PQ(pivotal quantity)로 사용하면 β_1에 대한 95% 신뢰구간은로

$$\begin{aligned} \hat{\beta}_1 \pm 2.571 \sqrt{\mathrm{MSE} / \sum (x_i - \bar{x})^2} &= 0.6 \pm 2.571 \sqrt{8/1080} \\ &= 0.6 \pm 0.2213 \end{aligned} \tag{6.4.8}$$

을 얻는다. (비고 : $0.025 = P(T_5 > 2.571) = P(T_5 < -2.571)$.)

다음, "$H_0: \beta_1 = \beta_{10}$"에 대한 검정통계량은 식 (6.4.7)의 β_1을 β_{10}로 대체한 것이다. 예를들어,

$$H_0: \beta_1 = 0, \quad H_a: \beta_1 \neq 0 \tag{6.4.9}$$

에 대해서 $\alpha = 5\%$ 로 검정하면

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{\mathrm{MSE}/\sum(x_i - \bar{x})^2}} = \frac{0.6}{\sqrt{8/1080}} = 7.746 > 2.571 \qquad (6.4.10)$$

이므로, 귀무가설 "$\beta_1 = 0$"을 기각한다.

위의 결과에 대한 해설은 다음과 같다. 첫째, <비고 5.5.5>에서 언급했듯이, 식 (6.4.8)의 "95% 신뢰구간이 0을 포함하지 않음"과 식 (6.4.9)의 "가설을 $\alpha = 5\%$ 로 기각함"은 동치이다. 둘째로, 식 (6.4.1)과 (6.4.2)에 의해서, SLR에서는 식 (6.4.9)의 가설이

$$H_0 : \text{Null Model} \quad H_a (\cup H_0) : CM$$

과 동치인데, 이는 다음과 같이 확인할 수 있다. $\alpha = 5\%$ 에 대해서 식 (6.4.3)은 $f = 60 > 6.61$인데, 이는 식 (6.4.10)을 제곱한 $t^2 = (7.746)^2 > (\pm 2.571)^2$과 일치한다.

통계 패키지는 ANOVA Table과 함께 (§6.4.1 참조) 다음과 같은 정보를 제공한다.

Parameter	Estimate	T for H:PARA=0	$PR>\|T\|$	STD Error of EST
Intercept	60	3.996	0.0104	15.014
X	0.6667	7.746	0.0006	0.0861

위의 표에서 "Parameter"는 β_0와 β_1을 의미하고, "Estimate"는 $\hat{\beta}_0$과 $\hat{\beta}_1$을 의미한다. 그리고 셋째 열에서 "7.746"은 바로 식 (6.4.10)의 t값이다. 이와 같이 "4.300"은 "$H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$"에 대한 t값이다 (§6.4.3 참조). 넷째 열은 "p-value"인데, "$PR>|T|$"에서 T에 절대값을 취한 것은 T-test가 TTT임을 의미하는 것이다. (비고 : 0.0006은 ANOVA Table의 $PR>F$값과 같음.) 마지막으로, 다섯째 열은 예를 들어 식 (6.4.10)의 분모인 $\sqrt{8/1080} \simeq 0.0861$인데, 이는 T-test에서 소음(noise)에 해당된다.

6.4.3 β_0에 대한 추론

식 (6.3.16)과 (6.4.5)에 의해서, 추정량 $\hat{\beta}_0$의 분포는

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x^2}}{\sum(x_i - \bar{x})^2})) \qquad (6.4.11)$$

인데, σ^2을 식 (6.3.15)의 MSE로 대체하면 (<비고 6.3.6> 참조)

$$T_{n-2}=\frac{\hat{\beta}_0-\beta_0}{\sqrt{\mathrm{MSE}\left(\frac{1}{n}+\frac{\overline{x^2}}{\sum(x_i-\overline{x})^2}\right)}}$$

를 얻는다 (식(6.4.7) 참조).

위의 식에서 분모를 계산하면

$$\sqrt{8\left(\frac{1}{7}-\frac{(174)^2}{1080}\right)}=15.014$$

를 얻는다. 이는 §6.4.2의 표에서 다섯째 열에 있는 값으로서 다음과 같이 구간추정과 검정에 쓰인다. 예를 들어, β_0에 대한 95% 신뢰구간은

$$\hat{\beta}_0\pm(2.571)(15.014)=60\pm38.6 \qquad (6.4.12)$$

이고, "H_0:$\beta_0=0$, H_a:$\beta_0\neq0$"에 대한 검정은

$$t=\frac{\hat{\beta}_0-0}{15.014}=\frac{60}{15.014}=3.996>2.571$$

이므로 $\alpha=5\%$에서 귀무가설 "$\beta_0=0$"를 기각한다.

t값인 3.996은 §6.4.2의 표에서 넷째 열에 있는 값이고, 이에 대한 p-value는 0.0104이다.

6.4.4 $Cov(\hat{\beta}_0,\hat{\beta}_1)$

식 (6.3.16)과 (6.4.5)에 의해서, 추정량 $\hat{\beta}_0$와 $\hat{\beta}_1$간의 공분산은

$$Cov(\hat{\beta}_0,\hat{\beta}_1)=-\sigma^2\overline{x}/\sum(x_i-\overline{x})^2 \qquad (6.4.13)$$

이므로, $\hat{\beta}_0$와 $\hat{\beta}_1$간의 상관계수는 (식 (2.13.6) 참조)

$$\rho=\frac{Cov(\hat{\beta}_0,\hat{\beta}_1)}{\sqrt{V(\hat{\beta}_0)\,V(\hat{\beta}_1)}}=\frac{-\overline{x}}{\sqrt{\overline{x^2}+\sum(x_i-\overline{x})^2/n}}\left(=\frac{-\overline{x}}{\sqrt{\sum x_i^2/n}}\right)=-0.9975$$

이다. (비고 : $V(\widehat{\beta_0})$와 $V(\widehat{\beta_1})$은 식 (6.4.11)과 (6.4.6) 참조.) 따라서, 예제에서는 $\widehat{\beta_0}$와 $\widehat{\beta_1}$ 간에 음의 상관관계가 있으며, 극단적인 경우인 $\rho = -1$에 가깝다. 이는, $(\overline{x}, \overline{y})$가 회귀직선 상에 있으므로 (<비고 6.1.2> 참조), 기울기가 증가하면 y-절편은 감소하기 때문이다 (단, $\overline{x}>0$일 때). 만약, $\overline{x}<0$이면 기울기가 증가함에 따라 y-절편도 같이 증가하므로 양의 상관관계로 바뀐다. 그리고, $\overline{x}=0$이면 기울기의 변화와 무관하게 y-절편은 항상 $(0, \overline{y})$이므로 $\rho = 0$이 된다.

간혹, β_0가 특별한 의미를 가지는 경우가 있는데, 이 경우에는 β_0에 대한 추정과 검정이 의미가 있다. 그러나, 일반적으로는 $\widehat{\beta_0}$가 $\widehat{\beta_1}$에 의해서 결정되다시피 하므로 β_0에 대한 추론은 상대적으로 중요시하지 않는다.

6.4.5 σ^2에 대한 추론

σ^2에 대한 점추정량으로는 MVUE인 MSE를 사용한다 (<비고 6.3.6> 참조). 반면에, σ^2에 대한 구간추정이나 검정은 별로 중요하게 여겨지지 않기 때문에 간단히 언급하면 신뢰구간은 §3.7.1에서와 같고 검정은 §4.4.3에서와 같은데, 다만 식 (2.15.11)의 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$대신에 SSE를 사용하기만 하면 된다.

6.4.6 CLM

지금까지 식 (6.3.16)에 관련된 추론을 했는데, 이제 식 (6.3.18)에 관련된 추론 (중에서 가장 대표적인 것)을 한다.

SLR 경우 식 (6.3.17)은 "$a_0\beta_0 + a_1\beta_1$"이다. 이에 "$a_0 = 1$, $a_1 = x$"를 대입하면

$$\mu_x \equiv \beta_0 + \beta_1 x \tag{6.4.14}$$

가 되는데, 이를 μ_x라 부르는 이유는 식 (6.1.1)에 의해서

$$Y_x \sim N(\mu_x, \sigma^2) \tag{6.4.15}$$

이기 때문이다.

먼저, 식 (6.1.1)과 식 (6.4.1)의 차이점을 지적한다. 식 (6.4.1)은 독립변수의 관찰치인 $x_1, \cdots, x_n$과 이에 대응하는 종속변수 $Y_1, \cdots, Y_n$만을 염두에 둔 것이다. 반면에, 식 (6.1.1)

에서는 독립변수 x가 (실제로 관찰된 $x_1,\cdots,x_n$뿐만 아니라) 모든 실수값을 가질 수 있는 변수(variable)로 취급되고 있다.

<비고 6.4.1> 식 (6.1.1)의 Y_x는 (일종의) 조건부 확률변수이다. 즉, Y_x는 독립변수의 값이 x일 때의 종속변수를 의미한다.

식 (6.4.14)의 μ_x가 x의 함수이므로, μ_x에 대한 점 추정량인

$$\hat{\mu}_x \equiv \hat{\beta}_0 + \hat{\beta}_1 x$$

도 x의 함수이고 또한 μ_x에 대한 신뢰구간도 x의 함수가 되는데, μ_x에 대한 신뢰구간을 흔히 CLM(confidence limit for the mean)이라 부른다.

$\hat{\mu}_x$의 분포는 식 (6.3.18)에 $a' = (a_0\ a_1) = (1\ x)$와 식 (6.4.5)를 대입하여 다음과 같이 얻는다.

$$\hat{\mu}_x \sim N\left(\mu_x, \sigma^2\left\{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right\}\right) \tag{6.4.16}$$

그리고, 종전과 같이 σ^2을 MSE로 대체하면 PQ로 사용할

$$T_{n-2} = \frac{\hat{\mu}_x - \mu_x}{\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right)}} \tag{6.4.17}$$

를 얻는다.

사용중인 예제에서 신뢰수준이 95%인 CLM은 다음과 같다.

$$\begin{aligned} \hat{\mu}_x \pm 2.571 &\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i-\bar{x})^2}\right)} \\ &= (60 + 0.6x) \pm 2.571\sqrt{8\left(\frac{1}{7} + \frac{(x-174)^2}{1080}\right)} \end{aligned} \tag{6.4.18}$$

<비고 6.4.2> $\mu_x \equiv \beta_0 + \beta_1 x$에 대한 신뢰구간인 CLM은 x의 함수인데, 신뢰구간의 폭은 $x = \bar{x}$일 때 최소이고 x가 $\bar{x}$에서 멀어질수록 증가한다.

예를 들어,

$$\begin{aligned}&(95\%\ \text{CLM when } x=168) = 172\pm 3.052\\&(95\%\ \text{CLM when } x=174) = 176\pm 2.749\\&(95\%\ \text{CLM when } x=180) = 180\pm 3.052\\&(95\%\ \text{CLM when } x=222) = 208\pm 10.971\end{aligned} \tag{6.4.19}$$

인데, 유의할 점은 다음과 같다. 첫째, $x=168$은 관찰치인 $x_3=168$과 일치하지만 나머지 x값들은 관찰치 $x_1,\cdots,x_7$과 다르다. 둘째로, 관찰치들의 범위인 "156~192"를 벗어나는 $x=222$ 경우에는 CLM의 폭이 (10.971로) 상당히 크다.

<비고 6.4.3> 식 (6.4.18)에 "$x=0$"을 대입하면 β_0에 대한 95% 신뢰구간인 식 (6.4.12)가 된다.

6.4.7 예측구간 (CLI)

갓 결혼한 남자의 키가 x라 하자. 만약, 아들이 태어난다면 아들이 성장했을 때 키는 얼마나 되겠는가? 아들의 키를 Y라 하면 <비고 6.4.1>에 의해서

$$Y=\beta_0+\beta_1 x+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2) \tag{6.4.19}$$

이다 (식 (6.1.1.) 참조). 즉, SLR에 근거한 아들의 키는 확률변수로서 분포는

$$Y\sim N(\beta_0+\beta_1 x,\sigma^2)$$

이다 (식 (6.4.15) 참조).

예측(prediction)도 일종의 추정이지만 차이점은 다음과 같다. 지금까지 점추정과 구간추정의 대상은 μ, σ^2, β_j 등의 모수였는데, 이들은 모르는(unknown)값들일 뿐 확률변수는 아니었다. (다만, 추정량이 확률변수인데, 이는 추정량이 확률변수 $Y_1,\cdots,Y_n$의 함수이기 때문이다.) 그러나, 예측에서는 예측의 대상인 Y 자체가 확률변수이다.

식 (6.4.19)에서 β_0와 β_1을 각각 식 (6.4.11)의 $\widehat{\beta_0}$과 식 (6.4.6)의 $\widehat{\beta_1}$으로 대체한 것을 $\widehat{Y}$라 하자. 즉,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon \tag{6.4.20}$$

인데, 이는 Y에 대한 점추정량인 셈이다. 이때 유의할 점은 식 (6.4.20)에서 ε이 $(\hat{\beta}_0 + \hat{\beta}_1 x)$와 독립이라는 점이다. β_0와 β_1에 대한 추정량인 $\hat{\beta}_0$와 $\hat{\beta}_1$는 확률변수 $Y_1, \cdots, Y_n$의 함수인데, $Y_1, \cdots, Y_n$이 서로 독립인 이유는 $\varepsilon_1, \cdots \varepsilon_n$이 *iid* 확률변수이기 때문이다 (<비고 6.3.1> 참조). 그런데, SLR 모형인 식 (6.1.1)에 의하면 ε_x는 표본과 관련된 $\varepsilon_1, \cdots \varepsilon_n$뿐만 아니라 표본과 무관한 ε_x들에 대해서도 *iid* 확률변수이다.

<비고 6.4.4> ε_x는 같은 x값에 대해서도 *iid*이다. 예를 들어, 같은 아버지의 여러 아들들의 키는 *iid* $N(\beta_0 + \beta_1 x, \sigma^2)$이다.

식 (6.4.20)에서, $(\hat{\beta}_0 + \hat{\beta}_1 x)$과 ε은 서로 독립이고 또한 정규분포를 따르므로 (비고 : 식 (6.4.16)이 $(\hat{\beta}_0 + \hat{\beta}_1 x)$의 분포임), $\hat{Y}$의 분포는 <비고 2.15.1>에 의해서

$$\hat{Y} \sim N\left(\beta_0 + \beta_1 x, \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right\}\right) \tag{6.4.21}$$

가 된다. 즉 $(\hat{\beta}_0 + \hat{\beta}_1 x)$의 분포인 식 (6.4.16)에서 분산만 σ^2만큼 증가시킨 것이 $\hat{Y}$의 분포이다. (비고 : $E(\varepsilon) = 0$이므로 평균은 증가하지 않음.)

이후 과정은 종전과 같다. 즉, 식 (6.4.21)의 σ^2을 MSE로 대체하면 PQ로 사용할

$$T_{n-2} = \frac{\hat{Y} - (\beta_0 + \beta_1 x)}{\sqrt{\mathrm{MSE}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)}}$$

를 얻는다. 따라서, 사용중인 예제에서 Y에 대한 95% 신뢰구간은 다음과 같다.

$$\hat{Y} \pm 2.571 \sqrt{\mathrm{MSE}\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)} \tag{6.4.22}$$

이제, Y에 대한 점추정량인 $\hat{Y}$를 점추정치 $\hat{y}$로 대체할 때가 되었다. 식 (6.4.20)에서 $Y_1, \cdots, Y_n$을 관찰치 $y_1, \cdots, y_n$으로 대체하면 $\hat{\beta}_0$와 $\hat{\beta}_1$은 추정량에서 추정치로 바뀐다. 그리고, 관행상 $E(\varepsilon) = 0$를 ε에 대한 추정치로 사용한다. 따라서, Y에 대한 점추정치는

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{6.4.23}$$

가 되어 결국 식 (6.4.18)의 $\hat{\mu}_x$와 일치한다.

<비고 6.4.5> $\hat{y}$이 바로 §6.1에서 예측치라 불렀던 것이다.

<비고 6.4.6> 식 (6.4.22)의 $\hat{Y}$을 식 (6.4.23)의 $\hat{y}$으로 대체한 것을 예측구간(prediction interval)이라 부른다. 또한, 식 (6.4.18)을 CLM이라 부르듯이, 예측구간을 CLI(confidence limit for an individual Y)라 부르기도 한다.

<비고 6.4.7> 흔히 CLM과 CLI를 혼동하는데, 가장 큰 이유는 점추정치가 같기 때문이다.

예를 들어,

$$(95\%\ \text{CLI when } x=168) = 172 \pm 7.887$$

$$(95\%\ \text{CLI when } x=174) = 176 \pm 7.774$$

$$(95\%\ \text{CLI when } x=180) = 180 \pm 7.887$$

$$(95\%\ \text{CLI when } x=222) = 208 \pm 13.162$$

인데, 유의할 점은 다음과 같다. 첫째, 식 (6.4.19)와 비교하면, 점추정치는 같지만 (<비고 6.4.7> 참조) 구간의 폭은 상당히 증가했다. 둘째로, $x=168$은 관찰치인 $x_3=168$과 일치하지만, 키가 x_3인 아버지의 아들의 키 y_3는 이미 관찰된 것인 반면에 키가 $x=168$인 아버지의 아들의 키 Y는 아직 관찰되지 않은 것이다.

§6.5 MLR의 분석

6.5.1 예제

§6.4에서는 SLR에 관해서 자세히 다루면서 결과들을 직접 계산하거나 유도하기도 하였다. 그러나, MLR은 SLR보다 복잡해서 간단한 예제에서나 직접 계산을 할 뿐, 일반적으로는 통계 패키지에 의존하게 된다. 사실, SLR을 자세히 다룬 이유도 패키지에 의한 MLR의 분석결과를 잘 이해하기 위한 것이다.

§6.2.4의 예제에 대한 패키지의 결과(output)는 다음과 같다.

Source of Variation	SS	자유도	MS	F	$PR>F$	R-SQUARE
Model	5.5429	2	2.7714	12.13	0.0762	0.9238
Error	0.4571	2	0.2286			
Total	6.0000	4				

Parameter	Estimate	T for H:PARA=0	$PR>\|T\|$	STD Error of EST
Intercept	0.5714	1.71	0.2285	0.3332
X_1	0.7000	4.63	0.0436	0.1512
X_2	0.2143	1.68	0.2355	0.1277

첫째, §6.2.4에서 계산했던 추정치 $\hat{\beta}=(4/7,\ 7/10,\ 3/14)$과 $SSE=0.4571$을 확인할 수 있다. 둘째로,

$$R^2=\frac{\mathrm{SSM}}{\mathrm{TSS}}\left(=\frac{\mathrm{SSE}_0-\mathrm{SSE}_{\mathrm{CM}}}{\mathrm{SSE}_0}\right)=\frac{5.5429}{6.0000}=0.9238 \qquad (6.5.1)$$

이다 (식 (6.4.4) 참조). 즉, 회귀모형이 전체 variation의 92.38%를 설명한다. 이때, "TSS=6"은 식 (6.3.14)에 의한 것이고 자유도는 "$n-1=4$"이다. 또한, 식 (6.3.15)에서 $\mathrm{SSE}_{\mathrm{CM}}=0.4571$이고 자유도는 $n-k-1=5-2-1=2$ 이다. 따라서 $\mathrm{SSM}=\mathrm{TSS}-\mathrm{SSE}_{\mathrm{CM}}=5.5429$이고 자유도는 $k-0=2$이다 (<비고 6.3.5> 참조).

셋째로, <비고 6.3.4>에 의한 "F-test for Model"의 결과는

$$F_{2,2}=\frac{\mathrm{SSM}/2}{\mathrm{TSS}/4}=\frac{2.7714}{0.2286}=12.13$$

인데 (식 (6.3.11)참조), p-value가 0.0762이므로 귀무가설 "$\beta_1=\beta_2=0$"을 $\alpha=5\%$ 로는 채택하고 $\alpha=10\%$ 로는 기각한다 (§4.6.3 참조).

넷째로, "$H_0:\beta_j=0$, $H_a:\beta_j\neq 0$"에 대한 T-test결과는 아래쪽 표에 있다. 예를 들어,

$$t=\frac{\widehat{\beta_1}-0}{(SDT\ Error\ of\ EST)}=\frac{0.7}{0.1512}=4.63 \qquad (6.5.2)$$

인데 (식 (6.4.10) 참조), p-value가 0.0436이므로 $\alpha=5\%$ 로 귀무가설 "$\beta_1=0$"를 기각한다. 반면에, 귀무가설 "$\beta_0=0$"와 "$\beta_2=0$"에 대한 p-value는 20%이상이므로 $\alpha=10\%$ 에서조차 귀무가설을 채택한다. 참고로 식 (6.5.2)에서 0.1512는 다음과 같이 얻을 수 있다. 추정량 $\widehat{\beta_1}$은 평균이 β_1인 정규분포를 따르는데, 식 (6.2.22)에 의해서 분산은 $\sigma^2/10$이다. 따라서, 분산의 추정치는 MSE/10 = 0.02286 이고, $\sqrt{0.02286}=0.1512$ 이다.

위의 결과들 외에도 CLM과 예측구간 (또는 CLI) 등 다양한 결과를 통계패키지로 얻을 수 있다.

6.5.2 예제 뒷처리

§6.5.1의 예제에서 $\alpha=5\%$ 로 "F-test for Model"에 대한 귀무가설 "$\beta_1=\beta_2=0$"을 채택했다. 그렇다면 과연 "$\beta_1=0$"이고 "$\beta_2=0$"인가? 반면에, β_1에 대한 T-test에서는 $\alpha=5\%$ 로 귀무가설 "$\beta_1=0$"를 기각했는데, 이는 과연 모순된 결과인가?

사실, MLR에서 가장 중요하고 또한 가장 어려운 주제는 소위 "Model Selection"이라는 것이다. Model Selection 이란 종속변수를 조금씩이나마 설명할 수 있는 모든 가능한 독립변수들 중에서 어떤 것은 취하고 어떤 것은 버릴 것인가를 결정하는 것이다. 즉, 모든 가능한 독립변수를 모두 포함시킨 모형을 CM이라 (하고 독립변수의 개수를 k라) 하면, $\{\beta_1,\cdots,\beta_k\}$ 중에서 하나 이상을 0으로 놓은 것이 RM인데 (§6.3.3 참조), RM들 중에서 가장 좋을 것을 선택하는 것이 Model Selection 이다.

<비고 6.5.1> RM의 총수는 (2^k-1)인데, 이는 k개의 독립변수 각각이 모형에 포함될 수도 있고 안될 수도 있기 때문이다. (단, CM을 제외하기 위해서 1을 뺌.)

예제에서는 $k=2$이므로 CM과 RM들은

$$
\begin{aligned}
&\text{CM} \ : Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i \\
&\text{RM}_1 : Y_i=\beta_0+\beta_1 x_{i1}+\varepsilon_i \\
&\text{RM}_2 : Y_i=\beta_0+\beta_2 x_{i2}+\varepsilon_i \\
&\text{RM}_3 : Y_i=\beta_0+\varepsilon_i
\end{aligned}
\tag{6.5.3}
$$

인데, ε_i의 분포는 모두 *iid* $N(0,\sigma^2)$이고 RM_3는 Null Model이라 부르던 것이다 (<비고 6.3.4> 참조).

유의수준 α를 5%로 책정하자. "F-test for Model"은

$$H_0 : \text{RM}_3,\ \text{H}_a : \text{CM}$$

이므로 (식 (6.3.12) 참조), H_0를 채택한다는 것은 CM과 비교했을 때 RM_3가 낫다는 뜻이다. 그렇지만, 아직 RM_3를 RM_1 또는 RM_2와 비교하지는 않았다. 반면에, β_1에 대한 T-test는

$$H_0 : \text{RM}_2,\ \text{H}_a : \text{CM}$$

이므로, H_0를 기각한다는 것은 CM이 RM_2보다 낫다는 뜻이다. 또한, β_2에 대한 T-test는

$$H_0 : \text{RM}_1,\ \text{H}_a : \text{CM}$$

이므로, H_0를 채택한다는 것은 RM_1이 CM보다 낫다는 뜻이다.

이상의 결과를 통합하면 ("$<$"는 선호도를 나타냄)

$$\text{RM}_2 < \text{CM} < \text{RM}_1, \text{RM}_3$$

이므로, 아직도 RM_1과 RM_3를 비교해 봐야한다. 가설을

$$H_0 : \text{RM}_3,\ \text{H}_a : \text{RM}_1$$

이라 하면 (식 (6.3.12) 참조), 이는 바로 독립변수 X_1 하나만 사용하는 SLR에서의 "F-test for Model" (인 동시에 β_1에 대한 T-test)이다. SLR에 대한 패키지의 결과는 다음과 같다.

Source of Variation	SS	자유도	MS	F	$PR>F$	R-SQUARE
Model	4.9	1	4.9	13.36	0.0354	0.81
Error	1.1	3	0.3			
Total	6.0	4				

Parameter	Estimate	T for H_0:PARA=0	$PR>\|T\|$	STD Error of EST
Intercept	1.0	3.69	0.0345	0.2708
X_1	0.7	3.66	0.0354	0.1915

p-value인 0.0354가 $\alpha = 0.05$보다 작으므로 귀무가설 "RM_3"를 기각하고 대립가설 "RM_1"을 채택한다. 따라서, RM_1을 가장 좋은 모형으로 선택하게 된다.

6.5.3 $Cov(\hat{\beta}_1, \hat{\beta}_2)$

사용중인 예제의 특징은 식 (6.2.22)의 $(X'X)^{-1}$에 "0"이 많이 있다는 점이다. 먼저 1행에 있는 0을 설명한다. $(X'X)^{-1}$에서 3행과 3열을 제거하면 독립변수 X_1 하나만 사용하는 SLR에서의 $(X'X)^{-1}$이 되는데, 이를 식 (6.4.5)와 비교하면 0이 발생한 이유가 "$\overline{X_1} = 0$"임을 알 수 있다. (비고 : 이에 대한 기하학적인 해석은 §6.4.4 참조.)

사실, 중요한 것은 3행에 있는 0인데, 이는 식 (6.3.16)에 의하면 "$Cov(\hat{\beta}_1, \hat{\beta}_2) = 0$을 의미한다. 그런데, 역시 식 (6.3.16)에 의하며 $\hat{\beta}_1$과 $\hat{\beta}_2$는 정규분포를 따르므로, "$Cov(\hat{\beta}_1, \hat{\beta}_2) = 0$은 "$\hat{\beta}_1$과 $\hat{\beta}_2$가 독립"임을 의미한다 (<비고 2.13.4> 참조).

<비고 6.5.2> 추정량 $\hat{\beta}_1$과 $\hat{\beta}_2$가 독립일 때 독립변수 X_1과 X_2가 독립이라고 표현하기로 한다.

추정량은 확률변수이므로 독립운운하는 것이 자연스럽다. 반면에, 독립변수는 확률변수가 아니라고 했으므로 독립운운하기가 어색한데, 이는 §6.6.6에서 논하기로 하자.

$\hat{\beta}_1$과 $\hat{\beta}_2$가 독립이든 X_1과 X_2가 독립이든 이에 따른 중요한 결과는 다음과 같다.

X_1만 사용하는 SLR에서 SSM은 4.9이고, X_2만 사용하는 SLR에서 SSM은 9/14인데, X_1과 X_2를 모두 사용하는 MLR에서 SSM은 (4.9+ 9/14)이다. 즉, Variation인 TSS=6 중에서 X_1 혼자서 설명하는 부분은 4.9이고, X_2 혼자서 설명하는 부분은 9/14인데, X_1과 X_2가 함께 설명하는 부분은 (4.9+ 9/14)이다.

이러한 결과는 당연히 그리고 항상 성립해야 되는 것으로 오해하기 쉽다. 그러나, 이러한 결과는 이상적인(?) 상황에서나 성립할 뿐, 일반적으로는 성립하지 않는다. 대체로, X_1과 X_2가 함께 설명하는 양이 따로따로 설명하는 양들을 합친 것보다 작다. 예를 들어, X_1은 아버지 키이고 X_2는 할아버지의 키라 하자. 종속변수인 아들의 키를 X_2보다는 X_1이 더 잘 설명하겠지만, X_2만으로도 어느정도 설명이 가능하다. 식 (6.4.4)에 의하면 X_1만으로 TSS의 92.31%를 설명한다. 그리고, 예를 들어, X_2만으로는 TSS의 60%를 설명한다고 하자. 그러나, X_1과 X_2를 모두 사용하는 MLR에서 TSS의 152.31%가 설명될 수는 없는 것이다.

<비고 6.5.3> 드물기는 하지만, X_1과 X_2가 함께 설명하는 양이 따로따로 설명하는 양들을 합친 것보다 클 경우도 있음 (문헌 [8] 참조).

6.5.4 Model Selection

§6.5.2의 예제에서는 $Cov(\widehat{\beta_1}, \widehat{\beta_2}) = 0$이라서 (§6.5.3 참조),

$$\mathrm{SSM}(X_1, X_2) = \mathrm{SSM}(X_1) + \mathrm{SSM}(X_2) \tag{6.5.5}$$

가 성립했고 따라서 Model Selection 과정이 비교적 간단했다. 그런데, 설사 식 (6.5.5)가 성립하지 않았다고 해도 $k=2$이기 때문에 모두 $2^k=4$개의 모형만 서로 비교하면 되었다 (식 (6.5.1) 참조). 그리고, 비교하는 횟수는 최대한 $\binom{4}{2}=6$회에 지나지 않는다. 그러나, 예를 들어 $k=10$이면, $2^{10}=1024$이고 $\binom{1024}{2}=523776$가 된다. 따라서, 효과적인 Selection 절차가 필요하다.

Model은 간단할수록 좋다. 즉, 독립변수의 수가 적을수록 좋다. 그러나, 최소한 모든 독립변수가 유의(significant)해야 한다. 즉, 모든 j에 대해서, β_j에대한 T-test의 p-value가 책정된 유의수준보다 작아야 된다. (비고: 모든 독립변수가 유의하면 "F-test for

Model"의 결과도 유의함.)

흔히, "max R^2"를 기준으로 사용하기도 하는데, 이는 잘못된 것이다. 물론, 그 이유는 어떤 독립변수를 추가하더라도 최소한 R^2가 감소하지는 않기 때문이다. 그러나, 동일한 개수의 독립변수를 사용하는 RM들 중에서는 R^2가 클수록 좋다. 예를 들어, 식 (6.5.3)에서 RM_1의 R^2가 RM_2의 R^2보다 크므로 RM_1이 RM_2보다 낫다.

다음, "min MSE"를 기준으로 사용하기도 하는데, 이는 최소한 "max R^2" 기준보다는 낫다. 별로 도움이 되지 않는 독립변수들을 제거하면 (SSM이 약간이나마 감소하므로) SSE가 약간 증가하기는 하지만 아울러 자유도도 증가하기 때문에 결과적으로 MSE가 감소하기도 하기 때문이다.

통계 패키지에 의한 방법은 크게 세가지가 있다. 첫째로, Forward 방법은 독립변수의 개수를 하나씩 증가시키는 방법이다. 예를 들어, 식 (6.5.3)에서는 RM_3로 시작한다. 1-단계에서는 RM_1과 RM_2 중에서 하나를 선택하는데, 예제에서는 RM_1을 선택하게 된다. (단, RM_1의 성능이 기준치 이상인 경우에 한함.) 2-단계에서는 독립변수를 하나 더 추가하는데, 예제에서는 CM이 유일하다. 그러나, 새로 추가된 X_2의 성능이 기준미달이므로 RM_1으로 낙착이 된다.

둘째로, Backward 방법은 CM으로 시작해서 독립변수를 하나씩 제거하는 방법이다. 예제에서는 1-단계에서 기준미달인 X_2를 제거한다. 그리고, 2-단계에서는 하나 남은 X_1의 제거여부를 결정한다.

사용중인 예제에서는 식 (6.5.5)가 성립하기 때문에 Forward와 Backward 방법의 결론은 동일하다. 그러나, 일반적으로는 그렇지 않다. 예를 들어, 처음에는 X_5가 힘을 발휘하다가 나중에 X_6가 등장하고 나서는 (X_5의) 힘이 약해지는 수가 있다. 또한 반대되는 현상도 가능하다 (<비고 6.5.3> 참조). 즉, 처음에는 X_5가 유의하지 않다가 나중에 X_6가 등장하면서 함께 큰 힘을 발휘하는 수도 있다. 이러한 점을 고려한 방법이 세 번째인 Stepwise 방법인데, 이는 Forward와 Backward 방법을 합친 것이라 할 수 있다. 즉, 일방동행이 아니라 양방향으로 왔다갔다 하면서 적절한 모형을 찾아내는 것이다.

6.5.5 선형모형의 범위

사용중인 예제는 다음과 같은 특징도 있다 (식 (6.2.21) 참조).

$$x_{i2} = x_{i1}^2 \ , \ i = 1, \cdots, n$$

즉, 두 번째 독립변수는 사실상 첫번째 독립변수를 제곱한 것이다. 그러니까, 회귀평면 (식 (6.2.24) 참조)

$$\hat{y} = 0.571 + 0.7 x_1 + 0.214 x_2$$

는 사실상 2-차원 상의 포물선

$$\hat{y} = 0.571 + 0.7 x_1 + 0.214 x_1^2$$

인 셈이다.

이와 같이, "선형"모형이라고 해서 반드시 "직선" 또는 "평면"에만 해당되는 것은 아니다. 문헌 [9]의 예제 11.10은 다음과 같다. SLR인

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ i = 1, \cdots, n \tag{6.5.6}$$

에서 $Y_i \equiv \ln W_i$, $\beta_0 = \ln \alpha_0$, $x = \ln I$, $\varepsilon = \ln \varepsilon'$이면 식 (6.5.6)은 사실상

$$W_i = \alpha_0 I_i^{\beta_1} \varepsilon_i', \ i = 1, \cdots, n \tag{6.5.7}$$

이다. 다만, 식 (6.5.6)에 대한 가정이 $\varepsilon_i \sim iid\ N(0, \sigma^2)$이므로, 식 (6.5.7)에서는 $\varepsilon_1', \cdots, \varepsilon_n'$이 (*iid* 확률변수이고) 대수 정규분포를 따르게 된다 (§2.9.3 참조).

식 (6.5.7)의 형태뿐만 아니라 어떤 형태라도 적절한 변환(transform)을 통해서 식 (6.5.6)의 형태가 (또는, 일반적으로 식 (6.3.1)의 형태가) 되기만 하면 이를 선형회귀모형으로 간주할 수 있다.

§6.6 상관관계 분석

6.6.1 서론: 독립변수도 확률변수?

SLR 대신에 상관분석(correlation analysis)으로 독립변수와 종속변수의 관계를 분석하기도 한다. 상관분석은 간편해서 중학교 과정에조차 등장하지만, 그 배경에 깔린 가정은 제법 복잡한 편이다.

상관관계란 "두개의 확률변수" 간의 선형종속성을 의미하는데, 이에 대한 표준화된 척도가 식 (2.13.6)의 상관계수이다. 따라서, 상관분석에서는 "독립변수도 확률변수"로 취급해야 된다.

SLR 대신에 상관분석을 할 때에는 독립변수 X와 종속변수 Y의 결합분포가 BVN(bivariate normal : 이변량 정규)분포라고 가정한다 (식 (6.3.16) 참조). 그런데, BVN 분포는 2장에서 다루지 않았으므로, 이 기회에 정식으로 다룬다. 또한 이 기회에, 지금까지 독립변수를 확률변수가 아니라고 해왔던 점에 대해서도 명확히 한다.

6.6.2 BVN 분포

지금까지 등장한 다변량(multivariate)분포는 §2.1.6의 다항분포와 §2.2.3의 다변량 초기하 분포인데, 이들은 모두 이산분포이다. 반면에, 연속분포에 대해서는 §6.3.4에서 $\{\widehat{\beta}_0, \cdots, \widehat{\beta}_k\}$의 결합분포가 MVN(다변량 정규)분포라고 언급만 했을 뿐, MVN분포에 대한 설명은 없었다. 그러나 MVN은 복잡하므로 BVN만 다룬다.

$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$이라 하고, 각각의 밀도함수를 $f_X(x)$, $f_Y(y)$라 하자. 그리고 X와 Y의 결합밀도함수를 $f(x,y)$라 하면, X와 Y가 독립인 경우에는 $f(x,y) = f_X(x) \cdot f_Y(y)$이다. 그러나 일반적으로는

$$f(x,y) = f_X(x) \cdot f_{Y|X}(y \mid x) \tag{6.6.1}$$

인데, $f_{Y|X}(y \mid x)$는 □$X = x$□라는 조건 하에서 Y의 (조건부)밀도함수이다 (<비고 2.14.1> 참조).

식 (6.6.1)에서 $f_X(x)$와 $f(x,y)$는 다음과 같다.

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma_X} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \tag{6.6.2}$$

$$f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y}e^{-\frac{Q}{2}} \tag{6.6.3}$$

$$\text{where}\quad Q=\frac{1}{1-\rho^2}\{(\frac{x-\mu_X}{\sigma_X})^2+(\frac{y-\mu_Y}{\sigma_Y})^2-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\}$$

식 (6.6.3)의 Q에 들어있는 ρ는 바로 X와 Y간의 상관계수인데, □ $\rho=0$ □일 때 □ $f(x,y)=f_X(x)\cdot f_Y(y)$ □가 성립함을 쉽게 알 수 있다 (<비고 2.13.4> 참조).

BVN분포를 따르는 모집단에서 추출한 크기가 n인 표본을 $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$이라 하고, 표본의 관찰치를 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$이라 하자. 그러면, <비고 2.7.1>에 의해서 LF는

$$L(\rho,\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2)=\prod_{i=1}^{n}f(x_i,y_i)$$

인데, 이로부터 ρ에 대한 MLE로

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \tag{6.6.4}$$

을 얻는다. (비고: $-1\le r\le 1$)

<비고 6.6.1> r은 표본상관계수라 불리는데, 식 (6.6.4)의 분모와 분자를 n으로 나누면 각각 식 (2.13.6)의 분모와 분자에 대응된다.

6.6.3 상관분석

§6.1의 예제에서 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$을 관찰된 표본이라 하면,

$$r=\frac{720}{\sqrt{1080}\sqrt{520}}=0.9608$$

을 얻는다. 이때 유의할 점은

$$r^2=(0.9608)^2=0.9231=R^2 \tag{6.6.5}$$

이다. 즉, SLR에서는 식 (6.6.4)를 제곱한 것이 식 (6.4.4)의 R^2와 일치한다 (계산은 생략

함).

또한, r은 회귀직선의 기울기인 식 (6.2.6)의 $\hat{\beta}_1$과 다음과 같은 관계가 있다.

$$r = \hat{\beta}_1 \frac{\sqrt{\sum(x_i - \overline{x})^2}}{\sqrt{\sum(y_i - \overline{y})^2}} \tag{6.6.6}$$

<비고 6.6.2> $\hat{\beta}_1$은 □$-1 \le r \le 1$□인 r을 x방향으로는 $\sqrt{\sum(x_i - \overline{x})^2/(n-1)}$ 만큼 늘리고, y방향으로는 $\sqrt{\sum(y_i - \overline{y})^2/(n-1)}$ 만큼 늘린 것과 같다. 즉, r은 일종의 표준화된 기울기인데, 이에 표본 표준편차의 비율을 곱하면 $\hat{\beta}_1$이 된다.

마지막으로, 식 (6.6.5)와 (6.6.6)에 의해서, 식 (6.4.10)의

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{MSE/\sum(x_i - \overline{x})^2}} = \frac{0.6}{\sqrt{8/1080}} = 7.746 \ > 2.571 \tag{6.6.7}$$

은 다음과 일치한다.

$$t = \frac{r - 0}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.9608}{\sqrt{0.0769/5}} = 7.746 \ > 2.571 \tag{6.6.8}$$

즉, □H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$□에 대한 검정과 □H_0: $\rho = 0$, H_a: $\rho \neq 0$□에 대한 검정은 일치한다. 그런데, 전자는 □F-test for model□과 일치하고 후자는 □X, Y의 독립성 검정□과 일치하므로(<비고 2.13.4> 참조), 결국 SLR에서는 위의 네가지가 모두 일치한다.

6.6.4 직교회귀

상관분석을 하기 위해서 X와 Y의 결합분포가 BVN분포라 가정하였다. 그리고, 관찰된 표본을 $\{(x_1, y_1), \cdots, (x_n, y_n)\}$이라 하였다. 그러면, 관찰된 표본에 가장 잘 들어맞는 직선은 무엇인가?

SLR에서 회귀직선은 <비고 6.1.1>에 따른 것인데, 그때 □수직방향 거리□의 제곱합만 따진 이유는 Y만 확률변수라고 가정했었기 때문이다. 그러나, 이제는 X와 Y가 대등한 확률변수로 취급되고 있으므로, □최단거리□의 제곱합을 따진다 (<비고 6.2.1> 참조).

최단거리 제곱합이 최소가 되게하는 직교(orthogonal)회귀직선을 구하는 방법은 선형

대수학과 관련이 있는데, 그 결과를 요약하면 다음과 같다 (문헌 [3] 참조). 첫째, 직교 회귀직선도 $(\overline{x}, \overline{y})$를 통과한다 (<비고 6.1.2> 참조). 따라서, 기울기만 구하면 되는데, 편의상 좌표축을 옮겨서 $(\overline{x}, \overline{y}) = (0,0)$가 되게 하자. 그리고,

$$W \equiv \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{vmatrix}$$

이라 하자. 둘째로, $W'W$의 고유치(eigen value)를 λ_1, λ_2라 하고 (단, $\lambda_1 > \lambda_2$), 대응하는 고유 벡터(eigen vector)를 각각 V_1, V_2라 하자. 그러면, V_1의 연장선이 직교 회귀직선이 되고, λ_2는 최단거리 제곱합이 된다.

예를 들어,

$$W' = \begin{pmatrix} -5.2 & -3 & 1.2 & 3 & -1.4 & 1 & 4.4 \\ -3.6 & -4 & -3.4 & -1 & 4.8 & 3 & 4.2 \end{pmatrix}$$

이면, $W'W = \begin{pmatrix} 68.8 & 38.4 \\ 38.4 & 91.2 \end{pmatrix}$ 로부터

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 120 \\ 40 \end{pmatrix}, \quad V_1 = \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix}, \quad V_2 = \begin{pmatrix} -0.8 \\ 0.6 \end{pmatrix}$$

을 얻으므로, 직교 회귀직선은 $y = 1.3\dot{3}\,x$이고 SSE$= 40$이다. 참고로, $\lambda_1 = 120$은 7개의 점으로부터 (V_2의 연장선인) $y = -0.75\,x$까지의 최단거리 제곱합이다. 또한, 직교 회귀직선에 대해서도 역학적인 해석이 가능하나 이를 생략한다 (§6.2.2 참조).

6.6.5 회귀모형의 재해석

§6.5 이전에는 독립변수를 확률변수가 아니라고 했는데, 이제와서 X를 확률변수라 하고 (X, Y)가 BVN분포를 따른다고 하면 이는 과연 모순인가?

관찰된 표본 $\{(x_1, y_1), \cdots, (x_n, y_n)\}$을 얻는다는 것은 확률변수 쌍(pair)인 (X, Y)를 독립적으로 n번 구현(realize)시킨다는 뜻이다. 즉, X와 Y를 동시에 구현시키는 것인데, 이러한 관점에서 상관분석과 직교회귀분석을 한 셈이다.

반면에, §6.5 이전에서의 관점은 다음과 같다. 한마디로, X를 먼저 구현시켜서 $\{x_1, \cdots, x_n\}$을 얻은 다음, $i = 1, \cdots, n$에 대해서 □ $X = x_i$□라는 조건 하에서 Y를 구현시키는 것이다. 이에 <비고 6.4.1>에서 암시했듯이, §6.5 이전에 등장한 종속변수는 □조건

부 확률변수□이다. 그리고, 이 조건부 확률변수의 밀도함수는 식 (6.6.1)의 $f_{Y|X}(y \mid x)$인데, 이 역시 정규분포를 따름을 보일 수 있다.

구체적으로, 식 (6.4.1)에서 □ Y_i□는 사실상 □ $Y \mid X = x_i$□이고, $(\beta_0 + \beta_1 x_i)$는 $E(Y \mid X = x_i)$이며, ε_i의 분산인 σ^2은 $V(Y \mid X = x_i)$이다. 그리고, 이러한 관계는 식 (6.3.1)의 MLR에 대해서도 성립한다. 단, MLR 에서는 $(X_1, \cdots, X_k, Y)$가 MVN분포를 따르고, $i = 1, \cdots, n$에 대해서 □ $X_1 = x_{i1}, \cdots, X_k = x_{ik}$□라는 조건 하에서 Y를 구현시킬 따름이다.

6.6.6 독립변수가 독립?

$(X_1, \cdots, X_k, Y)$가 MVN분포를 따른다고 하자. 그러면, 이상적인 경우는 $X_1, \cdots, X_k$가 서로 독립인 경우인데, 이때 식 (6.5.5)가 다음과 같이 확장된다.

$$SSM(X_1, \cdots, X_k) = \sum_{j=1}^{k} SSM(X_j)$$

$X_1, \cdots, X_k$는 서로 독립이더라도 각각은 Y와 (선형) 종속이다. (비고: MVN가정하에서는 비선형 종속은 존재하지 않으므로 따질 필요가 없음. <비고 2.13.4> 참조.) 그러나, Y와 종속인 독립변수를 모두 모형에 포함시키는 것은 아니다. 모형은 간단할수록 좋다고 했는데, 이는 유의(significant)한 독립변수만 모형에 포함시키는 것을 의미한다 (§6.5.4 참조). (만약에, Y와 독립인 X_j를 실수로 CM에 포함시켰더라도 $SSM(X_j) = 0$이므로 X_j는 Model Selection 과정에서 제거된다.)

그런데, 실제 문제에서는 $X_1, \cdots, X_k$가 서로 독립이 아니다. 예를 들어, Y가 아들의 키일 때, X_1, X_2, X_3는 각각 아버지, 할아버지, 어머니의 키일 수 있다. 이 경우 Y가 X_1에 종속이면 X_1은 다시 X_2에 종속이다. 그리고, 예를 들어 키가 큰 (작은) 사람끼리 결혼하는 경향이 있다면 X_1과 X_3도 종속이다.

독립변수끼리 독립이 아니더라도 Model Selection 과정에서 어느정도까지는 불필요한 독립변수를 제거할 수 있다. 예를 들어, X_1이 Y에 대해서 설명하지 못한 부분 중에서 X_2가 추가로 설명하는 부분의 크기를 따져서 X_2의 유의성을 판단할 수 있다. 이때 한가지 주의할 점은 다음과 같다. 예를 들어, 아버지의 키인 X_1과 어머니의 키인 X_3간의 상관

계수가 1에 가깝다고 하자. 이 경우, X_1이 (또는, X_3가) 먼저 Y를 설명하고 나서 X_3가 (또는, X_1이) 추가로 설명하는 부분의 크기는 미미하므로 X_3는 (또는, X_1은) 당연히 모형에서 제거되어야 한다. 그러나, 이때 기술적인 (technical) 문제가 발생할 수 있다. 첫째, 극단적으로 X_1과 X_3간의 상관계수가 ± 1이면 식 (6.2.15)의 X의 rank가 하나 부족해서 $(X'X)^{-1}$가 존재하지 않는다. 둘째로, X_1과 X_3간의 상관계수가 ± 1은 아니더라도 ± 1에 가까우면 MSE가 상당히 커진다. 그런데 MSE는 회귀분석에 등장하는 모든 검정통계량에서 소음(noise)에 해당되는 분모에 포함되어 있으므로 MSE가 커지면 모든 검정의 검정력이 약해진다. 이러한 현상을 다중공선성(multicolinearity)이라 하는데, 이를 방지하기 위해서는 X_1과 X_3중에서 하나를 처음부터 CM에서 제거해야 된다. (예를 들어, X_1과 X_3간의 표본 상관계수가 ± 1에 가까우면 하나를 제거하는데, 물론 Y를 조금이나마 잘 설명하는 것을 남긴다.)

참고로, §6.5.2의 예제에서는 $X_2 = X_1^{\,2}$이므로 X_1과 X_2가 (최소한 비선형적으로는) 종속인데, X_1의 관찰치를 원점에 대해서 대칭이 되도록 $(-2,-1,0,1,2)$라 했기 때문에 결과적으로 선형적으로는 X_2와 독립이 되었다. 만약, (X_1, X_2, Y)가 MVN분포를 따른다는 가정이 합당하다면 선형독립은 독립을 의미한다 (<비고 6.5.2> 참조). 그러나 $(-2,-1,0,1,2)$는 정규분포를 따르는 X_1의 관찰치라 하기 어렵고, 오히려 인위적으로 설정한 값이 분명하다. 따라서, 예제에서는 X_1과 X_2가 확률변수라하기 어렵다.

회귀모형에서 독립변수는 확률변수일 수도 있고 아닐 수도 있는데, 특히 인위적으로 제어(control)할 수 있는 독립변수는 확률변수라 하기 어렵다.

사실 독립변수가 확률변수인지 아닌지에 대한 논의는 5장의 ANOVA로 거슬러 올라간다. §5.5.2 에서 Y_{ij}를 젖소 j에게 사료 i를 먹일 때의 우유생산량이라 했다. 그런데, 많은 젖소 중에서 J마리를 (무작위로) 뽑았다면, 식 (5.5.4)에서 β_j는 확률변수가 되어야 마땅하다. 또한, 여러 종류의 사료 중에서 I종류를 뽑았다면 식 (5.5.4)의 τ_i까지도 확률변수가 되는데, β_j와 τ_i가 확률변수가 되면 분석방법이 복잡해지므로 편의상 이들을 상수로 취급한 것이다. (비고: 자세한 내용은 실험계획법 교재 참조.)