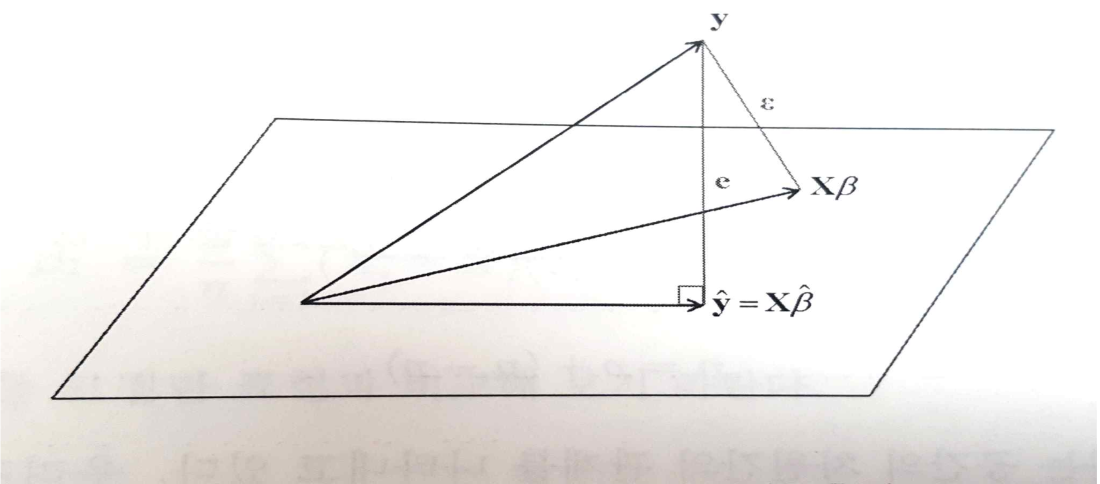
김은영

제3장 중선형회귀모형

3.8 기타논제

3.8.3 최소제곱법의 기하학적 의미

최소제곱법을 이용하여 회귀계수를 추정하는 과정은 기하학적으로 더욱 쉽고 명료하게 설명될 수 있다.



(그림 3.5) 추정공간과 최소제곱추정

중회귀모형 $y = X\beta + \epsilon$ 여기서, $X_{(n \times p)}$

■ 반응변수의 관측값 벡터 $y = [y_1 \cdots y_n]^t$: n차원 공간의 한 점

■ $X\beta = [X_0 \quad X_1 \quad \cdots \quad X_{p-1}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$

■ 만약 X의 계수(rank)가 p라면 $X\beta$는 β의 값이 변함에 따라 n차원 공간에서 p개의 선형독립인 벡터 $[X_0 \quad X_1 \quad \cdots \quad X_{p-1}]$에 의해 생성(span)되는 p차원 평면[→추정공간(estimation space)]을 형성하게 된다.

■ 오차벡터 $\epsilon = [\epsilon_1 \quad \epsilon_2 \quad \cdots \quad \epsilon_n]^t$

: 이 추상공간 위의 한 점과 벡터 y와의 차이를 나타내는 벡터

■ 최소제곱법: 오차제곱합[$S=\epsilon^t\epsilon=(y-X\beta)^t(y-X\beta)$]을 최소화하는 β의 값을 찾는 것
→ 벡터 y에서 가장 가까운 추정공간 위의 한 점을 찾는 것
→ 이 점은 벡터 y를 추정공간에 정사영(orthogonal projection)시켜 얻을 수 있으며, 이 점을 $\hat{y}=X\hat{\beta}$라 하면 여기에 해당되는 벡터 $\hat{\beta}$가 β의 최소제곱추정치가 된다.

■ 잔차벡터 $e=y-X\hat{\beta}$ 는 추정공간 위의 모든 벡터들과 직교한다.

X의 모든 열벡터에 대해 $X_j^t e=X_j^t(y-X\hat{\beta})=0 \qquad (j=0,1,\cdots,p-1)$

→ $X^t e=X^t y-X^t X\hat{\beta}=0 \quad \Rightarrow$ (정규방정식) $X^t X\hat{\beta}=X^t y$

■ 회귀의 분산분석에 사용되는 세 가지 제곱합의 분할은 직각삼각형에 대한 피타고라스 정리를 사용하여 얻어진다.

- 잔차벡터 e와 $\hat{y}$이 직교 → $\|y\|^2=\|\hat{y}\|^2+\|e\|^2$
- 벡터의 내적으로 표현 → $y^t y=\hat{y}^t\hat{y}+(y-X\hat{\beta})^t(y-X\hat{\beta}) \;\leftarrow\; SST=SSR+SSE$
- 자유도 간의 관계식 $n=p+(n-p)$: 각 벡터가 위치하는 공간의 차원 간의 관계

3.8.4 다변량 정규분포와의 관계

$Z\sim N_n(\mu,\ \Sigma)$

$Z=\begin{bmatrix}Z_1 & Z_2\end{bmatrix}^t,\ \mu=\begin{bmatrix}\mu_1 & \mu_2\end{bmatrix},\ \Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}$

→ $Z_1|Z_2\sim N_r\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}(Z_2-\mu_2),\ \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

$y=\beta_0+\beta_1X_1+\cdots+\beta_{p-1}X_{p-1}+\epsilon$

$$
\begin{aligned}
E(Y|X)&=\beta_0+\beta_1X_1+\cdots+\beta_{p-1}X_{p-1}\\
&=\beta_0+\begin{bmatrix}\beta_1\cdots\beta_{p-1}\end{bmatrix}^t X \qquad \text{여기서, } X=\begin{bmatrix}X_1 & \cdots & X_{p-1}\end{bmatrix}\\
&=E(Y)+Cov(Y,X)Cov(X)^{-1}[X-E(X)] \quad \leftarrow \text{다변량 정규분포의 조건부 기댓값}\\
&=E(Y)-Cov(Y,X)Cov(X)^{-1}E(X)+Cov(Y,X)Cov(X)^{-1}X
\end{aligned}
$$

$\Rightarrow \beta_0=E(Y)-Cov(Y,X)Cov(X)^{-1}E(X)$

$\begin{bmatrix}\beta_1\cdots\beta_{p-1}\end{bmatrix}^t=Cov(Y,X)Cov(X)^{-1}$

$p=2$인 단순선형회귀모형의 경우를 생각해 보면

모집단	표본
$\beta_0 = E(Y) - \frac{Cov(Y,X)E(X)}{Var(X)}$	$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$
$\beta_1 = \frac{Cov(Y,X)}{Var(X)}$	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$