稜線回歸模型에서의 變數의 正規化

鄭 基 俊*

<目 次>

I. 序 論

稜線回歸模型(ridge regression model)은 1970年 호얼과 케나드(1970)에 의해 개발된 이래 計量經濟學의 중요한 分析技法으로서 발전되어 왔다. 이 模型이 널리 관심을 끌게 된 것은 특히 經濟時系列資料의 分析에 있어서 불가피하게 직면하게 되는 共線形性(multicollinearity)의 문제를 이 模型으로 해결할 수 있으리라는 희망 때문이었고, 실제로 이 모형은 그 희망을 어느 정도 충족시켜 줄 수 있다는 것이 밝혀졌다. 그리하여 이 모형은 현재 상당히 널리 경제분석에 응용되고 있다.

이 능선회귀분석에서는 그 技法의 특징상 變數 특히 說明變數의 正規化가 반드시 요구된다. 이는 기본적으로 變數들을 無名數化하여, 상호간의 크기를 직접 비교할 수 있도록 한다는 요구로부터 나오는 것이다. 그러나 아래의 설명에서 볼 수 있는 바와 같이, 旣存文獻에서 소개되는 正規化의 方法은 적어도 理論的으로 엄정하지 못한 측면을 가지고 있다. 따라서 理論的 엄정성을 期하기 위해서는 旣存의 正規化方法에 대한 반성과 代案의 제시가 필요하다. 本論文에서는 이 과제에 대한 해결책을 모색해 보고자 한다.

또 旣存文獻에서는 모형 속에 언제나 상수항이 포함되는 것을 전제로 하고 논의가 전개되어 왔다. 기존문헌에서는 상수항이 포함되지 않은 모형에 대해서 명시적인 언급이 없이

* 本硏究所 硏究員, 서울大學校 經濟學科 敎授.

상수항이 있는 경우와 같이 다룰 수 있다고 보는 것 같으나, 사실은 그렇지 않다는 것을 본 논문은 밝히려 한다. 그리하여 상수항이 없는 경우에 사용할 수 있는 正規化의 방법을 제시하고, 이를 평가하고자 한다.

II. 模　型

통상적인 의미에서의 標準線形模型은 k개의 설명변수와 n개의 관측치를 가진 선형모형으로서 다음과 같이 規定된다.

(2.1) $y=X\beta+\varepsilon,\ \ \varepsilon\sim(0,\ \sigma^2 I)$

여기서 X는 $n\times k$ 설명변수 관측치 行列이며, 그 첫째 列은 常數項 변수의 벡터로서, 모든 元素가 1인 $n\times 1$ 벡터이다. 이 벡터를 1로 나타내기로 하자. 그러면 行列 X는 다음과 같이 分割된다. 즉,

(2.2) $X=[1\vdots x_2\ x_3\ \cdots\ x_k]=[1\vdots X_2]$

여기서 X_2는 상수항변수를 제외한 $k-1$개의 설명변수의 관측치로 된 $n\times(k-1)$ 行列이다.

稜線回歸에서 行列 X를 X^*로 正規化한다고 할 때, 그 의미는 行列 $X^{*\prime}X^*$가 X의 標本相關係數行列이 되도록 X^*를 規定하는 것을 말한다. 이를 좀더 구체적으로 설명하기 위하여, X의 元素를 x_{ij}, X^*의 元素를 x_{ij}^*로 표기하기로 하자($i=1,2,\cdots,n;\ j=1,2,\cdots,k$). 그러면 x_{ij}와 x_{ij}^* 간에는 다음 관계가 성립하여야 한다.

(2.3) $x_{ij}^*=(x_{ij}-\bar{x}_j)/s_j^*$

단, $\bar{x}_j$는 j번째 설명변수의 표본평균이고, S_j는 $s_{ij}-\bar{x}_j$의 平方合의 陽의 平方根으로, 각각 다음과 같이 정의된다. 즉,

(2.4) $\bar{x}_j=\sum_{i=1}^{n}x_{ij}/n$

(2.5) $s_j^*=\sqrt{\Sigma(x_{ij}-\bar{x}_j)^2}$

여기서 $n\times n$ 行列 A를

(2.6) $A=I-\frac{1}{n}11'$

로 정의하면, Ax_j의 i번째 元素가 바로 $x_{ij}-\bar{x}_j$임을 쉽게 확인할 수 있다. 따라서 $AX_2S_2^{*-1}$는 X_2의 正規化된 형태 즉 X_2^*임을 쉽게 확인할 수 있다. 즉,

(2.7) $X_2^*=AX_2S_2^{*-1}$

단

$$(2.8)\quad S_2^*\equiv\begin{bmatrix} s_2^* & & \bigcirc \\ & s_3^* \cdots & \\ \bigcirc & & s_k^* \end{bmatrix}$$

그러나 x_1 즉 1는 위의 방법으로 正規化되지 않는다. 왜냐하면 $A1=0$, $S_1=0$로 되기 때문이다.

以上의 正規化에 관한 설명은 模型 (2.1)에서 行列 X만을 分離시킨 상태에서의 설명이다. 그러면 模型 (2.1) 자체를 그대로 놓은 채 X를 正規化할 때, 이 모형은 어떻게 變形되는가를 보기로 하자. 正規化를 위해서는 우선 行列 A를 그 模型의 양변에 앞곱해야 한다. 그러면 그 모형은 다음과 같이 變形된다. 즉,

$$(2.9)\quad Ay=AX\beta+A\varepsilon,\quad A\varepsilon\sim(0,\ \sigma^2A)$$

여기서 $A\varepsilon$의 分散行列을 σ^2A로 쓴 것은,

$$(2.10)\quad A'=A,\quad A^2=A$$

라는 성질을 이용한 것이다. 그리고 이 變形된 模型의 右邊의 첫째 項을 고쳐 쓰면 다음과 같다.

$$(2.11)\quad AX\beta=A[1\,\vdots\,X_2]\beta=[0\,\vdots\,AX_2]\begin{bmatrix}\beta_1\\ \cdots\\ \beta_2\end{bmatrix}=AX_2\beta_2$$

여기서 β_1은 β의 원소중 常數項係數이며, β_2는 나머지 $k-1$개의 係數로 된 $(k-1)\times1$벡터이다. 즉 이 變形과정에서 常數項이 사라진다.

正規化를 完結하려면 AX_2의 뒤에 S_2^{*-1}를 곱해야 한다. 그런데 S_2^*는 非特異行列이므로, $S_2^{*-1}S_2^*=I$의 관계가 성립하여, 따라서 正規化의 작업은 다음과 같이 진행된다. 즉,

$$(2.12)\quad AX_2\beta_2=AX_2(S_2^{*-1}S_2^*)\beta_2=(AX_2S_2^{*-1})(S_2^*\beta_2)=X_2^*\beta_2^*$$

단

$$(2.13)\quad \beta_2^*\equiv S_2^*\beta_2$$

로 정의된다. 그리하여 식 (2.9)에 (2.11) 및 (2.12)의 관계를 代入하면, 표준선형모형 (2.1)의 正規化된 形態를 얻는다. 즉

$$(2.14)\quad Ay=X_2^*\beta_2^*+A\varepsilon,\quad A\varepsilon\sim(0,\ \sigma^2A)$$

이것이 稜線回歸의 기본모형이 되어야 하는 것이다.

여기서 X_2^*가 과연 正規化되었는가를 알아보기 위하여 $X_2^{*\prime}X_2^*$를 평가해 보면 다음과 같다. 즉 식 (2.7)에 의해서,

$X_2^{*\prime}X_2^{*}=S_2^{*-1}X_2'AX_2S_2^{*-1}$

그런데 여기서 $X_2'AX_2$의 (j,k) 번째 원소는 $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{jk}-\bar{x}_k)$이므로, 이를 s_js_k로 나눈 것이 $X_2^{*\prime}X_2^{*}$의 (j,k)번째 원소가 되며 따라서 이는 x_j와 x_k간의 상관계수가 된다.

III. 正規化된 模型의 문제점

表記의 방법은 다를 수 있어도 旣存文獻에서 볼 수 있는 正規化된 稜線回歸模型은 기본적으로 식 (2.14)로 주어진 모형이다. (박성현〔1981〕, 사와〔1979〕, 비노드〔1978〕) 편의상 이를 다시 써 보자.

(3.1) $Ay=X_2^{*}\beta_2^{*}+A\varepsilon,\ A\varepsilon\sim(0,\ \sigma^2A)$

그러나 式 (3.1)을 基本模型으로 삼는데는 적어도 두개의 문제점 내지 불만족스러운 점이 있다. 첫째, 교란항이 $A\varepsilon$의 형태로 도입된다는 사실이다. 표준모형 (2.1)의 가정대로 ε의 分散行列을 σ^2I라고 하면 $A\varepsilon$의 分散行列은 σ^2A일 수 밖에 없다. 그리고 A의 位數는 $n-1$이므로, $A\varepsilon$의 分散行列의 位數도 $n-1$을 넘을 수 없다. (이는 ε의 分散行列이 어떤 형태를 취하는 경우라도 妥當하다) 그럼에도 불구하고 稜線回歸에 관한 많은 文獻은, (結果的으로 $A\varepsilon$로 될 수 밖에 없는) 교란항의 分散行列을 σ^2I로 "假定"하고 있는데, 이는 위의 논의에서 볼 때, 있을 수 없는 假定임이 분명하다.

두번째의 문제점은 常數項이 탈락되어버리는 점이다. 이처럼 상수항이 탈락되어 설명변수의 관측치 행렬 X_2^{*}가 $n\times(k-1)$ 行列이 되기 때문에, Ay가 아니라 y에 관한 논의를 할 필요가 있을 때는 또 하나의 번거로운 절차를 거쳐야 한다는 문제점이 있는 것이다.

IV. 문제점의 解決策

먼저

(4.1) $Ay=y-1\bar{Y},\ A\varepsilon=\varepsilon-1\bar{\varepsilon}$

임을 고려하면서 正規化된 모형 (3.1)을 다음과 같이 變形해 보자. 즉,

(4.2) $y=1(\bar{Y}-\bar{\varepsilon})+X_2^{*}\beta_2^{*}+\varepsilon$

여기서 $\bar{Y}$와 $\bar{\varepsilon}$는 平均 즉, $\bar{Y}=\frac{1}{n}1'y,\ \bar{\varepsilon}=\frac{1}{n}1'\varepsilon$을 의미한다. 또 $\bar{Y}-\bar{\varepsilon}$의 의미를 알아보기 위하여 원래의 모형 (2.1)의 양변에 $\frac{1}{n}1'$을 앞곱하면 다음 결과를 얻는다. 즉,

(4.3) $\bar{Y}=\bar{X}\beta+\bar{\varepsilon}$

단 $1\times k$ 벡터 $\bar{X}$는 $\frac{1}{n}1'X$로서,

(4.4) $\bar{X}=\frac{1}{n}1'X=\frac{1}{n}1'[1 \vdots X_2]=[1 \vdots \bar{X}_2]$

로 쓸 수 있고, $\bar{X}_2$는 설명변수의 평균들로 구성되는 $1\times(k-1)$ 벡터이다. 즉,

(4.5) $\bar{X}_2=[\bar{x}_2,\ \bar{x}_3,\ \cdots,\ \bar{x}_k]$

이제 식 (4.3) 및 (4.4)를 이용하여 $\bar{Y}-\bar{\varepsilon}$를 다음과 같이 표현할 수 있다. 즉,

(4.6) $\bar{Y}-\bar{\varepsilon}=\bar{X}\beta=[1 \vdots \bar{X}_2]\begin{bmatrix}\beta_1\\ \cdots\\ \beta_2\end{bmatrix}=\beta_1+\bar{X}_2\beta_2$

이 식에서 알 수 있는 중요한 사실은, $\bar{Y}$와 ε가 確率變數임에도 불구하고, 그 差인 $\bar{Y}-\bar{\varepsilon}$는 確率變數가 아니라는 점이다. 식 (4.6)을 이용하여 模型 (4.2)를 다시 쓰면 다음과 같다.

(4.7) $y=1(\beta_1+\bar{X}_2\beta_2)+X_2^*\beta_2^*+\varepsilon$

이 식의 우변의 첫째 項은 常數項벡터이다. 그러므로 이 項을 $x_1^{*\prime}\beta_1^*$의 형태로 쓰고, 이때의 x_1^*가 $x_1^{*\prime}x_1^*=1$이라는 조건을 充足한다면, 그것은 우리가 바라는 模型이 될 수 있을 것이다. 우리는 이 조건을 충족하는 x_1^*와 β_1^*가

(4.8) $x_1^*=\frac{1}{\sqrt{n}}1,\ \beta_1^*=\sqrt{n}(\beta_1+\bar{X}_2\beta_2)$

임을 쉽게 확인할 수 있다. 그리하여

(4.9) $X^*=[x_1^* \vdots X_2^*],\ \beta^*=[\beta_1^* \vdots \beta_2^{*\prime}]'$

이라 놓으면, 식 (4.8) 및 (4.7)에서 우리는 다음 식을 얻는다. 즉,

(4.10) $y=X^*\beta^*+\varepsilon,\ \varepsilon\sim(0,\ \sigma^2I)$

이 모형은 분명히 제 3 절에서 지적된 두가지 문제점이 모두 제거된 正規化模型이다. 즉 이제 교란항은 分散行列이 σ^2I인 ε이며, X^*는 $n\times k$行列로서, 常數項 변수가 어떤 형태로든지 포함되어 있다.

V. 새로운 正規化模型의 의미

새로운 正規化模型 (4.10)이 바람직한 성질들을 가지고 있는지 알아보기 위하여, β^*와 X^*의 의미를 分析해보자. 식 (2.13)과 식 (4.8)에 의하면, β^*는 다음과 같이 정리될 수 있다. 즉,

(5.1) $\beta^*=\begin{bmatrix}\beta_1^*\\ \beta_2^*\end{bmatrix}=\begin{bmatrix}\sqrt{n}\beta_1+\sqrt{n}\bar{X}_2\beta_2\\ S_2^*\beta_2\end{bmatrix}=\begin{bmatrix}\sqrt{n} & \sqrt{n}\bar{X}_2\\ 0 & S_2^*\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$

따라서 行列 S^*를,

$$(5.2)\quad S^*=\begin{bmatrix}\sqrt{n} & \sqrt{n}\bar{X}_2\\ 0 & S_2^*\end{bmatrix}$$

로 정의하면, β^*는,

$$(5.3)\quad \beta^*=S^*\beta$$

로 나타낼 수 있다. 그리고 行列 S^*는 비특이행렬임을 쉽게 확인할 수 있다. 따라서 β^*와 β는 1대1의 대응관계를 가진다.

다음은 X^*를 분석하기 위하여, 식 (2.7)과 식 (4.8)을 이용하면, X^*는 다음과 같이 정리될 수 있다. 즉,

$$(5.4)\quad X^*=[x_1^*\vdots X_2^*]=\left[\frac{1}{\sqrt{n}}1\vdots AX_2S_2^{*-1}\right]$$

그런데 $AX_2S_2^{*-1}$는 또 다음과 같이 변형될 수 있다. 즉,

$$(5.5)\quad AX_2S_2^{*-1}=(I-\frac{1}{n}11')X_2S_2^{*-1}=X_2S_2^{*-1}-1\left(\frac{1}{n}1'X_2\right)S_2^{*-1}=X_2S_2^{*-1}-1\bar{X}_2S_{2*}{}^{-1}$$

$$=[1\vdots X_2]\begin{bmatrix}-\bar{X}_2S_2^{*-1}\\ S_2^{*-1}\end{bmatrix}$$

이를 식 (5.4)에 대입하면,

$$(5.6)\quad X^*=[1\vdots X_2]\begin{bmatrix}\frac{1}{\sqrt{n}} & -\bar{X}_2S_2^{*-1}\\ 0 & S_2^{*-1}\end{bmatrix}$$

를 얻는다. 그리고 이 식의 우변의 行列은 바로 S^{*-1}임을 쉽게 확인할 수 있다. 즉,

$$(5.7)\quad S^{*-1}=\begin{bmatrix}\frac{1}{\sqrt{n}} & -\bar{X}_2S_2^{*-1}\\ 0 & S_2^{*-1}\end{bmatrix}$$

따라서 식 (5.6)은 다음과 같이 고쳐 쓸 수 있다. 즉,

$$(5.8)\quad X^*=XS^{*-1}$$

이것이 X와 X^* 사이의 관계를 나타내는 식이다. 그리고

$$(5.9)\quad x_1^{*\prime}X_2^*=\frac{1}{\sqrt{n}}1'AX_2S_2^{*-1}=\frac{1}{\sqrt{n}}0'X_2S_2^{*-1}$$

이므로, $X^{*\prime}X^*$는

$$(5.10)\quad X^{*\prime}X^*=\begin{bmatrix}x_1^{*\prime}x_1^* & 0'\\ 0 & X_2^{*\prime}X_2^*\end{bmatrix}=\begin{bmatrix}1 & 0'\\ 0 & X_2^{*\prime}X_2^*\end{bmatrix}$$

로 分割表現될 수 있고, 이것은 X^*의 正規化의 의미를 나타내 주는 것으로 볼 수 있다. 즉 $X_2^{*\prime}X_2^*$가 상관계수행렬이고, $x_1^{*\prime}X_2^*$는 1과 X_2간의 상관계수행렬로 해석한다면, $X^{*\prime}X^*$는 대각원소가 모두 1이고, 어떤 의미에서 相關係數行列로 볼 수 있는 것이다.

VI. 새로운 正規化模型의 稜線回歸計算

새로운 正規化模型 (4.10)을 써서 稜線回歸計算을 해 보자. β^*의 능선회귀계수를 $b^*(\alpha)$라 하면, 이는 다음과 같이 정의된다. 즉,

(6.1) $b^*(\alpha)=(X^{*\prime}X^*+\alpha I_k)^{-1}X^{*\prime}y \quad \alpha\geqq 0$

그리고 식 (5.10)으로 제시된 $X^{*\prime}X^*$의 分割을 이용하여, 식 (6.1)을 分割表現하여 보자. 즉,

(6.2) $$b^*(\alpha)=\begin{bmatrix} b_1^*(\alpha) \\ b_2^*(\alpha) \end{bmatrix}=\begin{bmatrix} 1+\alpha & 0' \\ 0 & X_2^{*\prime}X_2^*+\alpha I_{k-1} \end{bmatrix}^{-1}\begin{bmatrix} x_1^{*\prime}y \\ X_2^{*\prime}y \end{bmatrix}$$

즉

(6.3) $$b_1^*(\alpha)=\frac{1}{1+\alpha}\,x_1^{*\prime}y=\frac{\sqrt{n}}{1+\alpha}\bar{Y}$$

(6.4) $$b_2^*(\alpha)=(X_2^{*\prime}X_2^*+\alpha I_{k-1})^{-1}X_2^{*\prime}y$$

로 된다.

식 (6.3)에 의하면 $\alpha=0$일 때, 즉 통상최소자승의 경우 $b_1^*(0)=\sqrt{n}\bar{Y}$이며, 이것은 β_1^*의 不偏推定量이다. 그러므로 $b_1^*(\alpha)$는 $1/(1+\alpha)$에 비례하여 β_1^*를 低評價한다. 한편 식 (6.4)에 의하면 $b_2^*(\alpha)$는 y를 X_2^*에 능선회귀시킬 때의 回歸係數라고 해석할 수 있다. 그런데 X_2^*의 定義式 (2.7)과 行列 A의 성질로부터,

(6.5) $$X_2^{*\prime}y=S_2^{*-1}X_2'Ay=S_2^{*-1}X_2'AAy=X_2^{*\prime}Ay$$

로 쓸 수 있으므로, 식 (6.4)는,

(6.6) $$b_2^*(\alpha)=(X_2^{*\prime}X_2^*+\alpha I_{k-1})^{-1}X_2^{*\prime}Ay$$

로 고쳐쓸 수 있다. 식 (6.6)은 $b_2^*(\alpha)$가 Ay를 X_2^*에 능선회귀시킬 때의 회귀계수임을 의미한다. 즉, 통상적 正規化模型 (2.14)에 능선회귀방법을 적용할 때의 회귀계수이다. 이는 극히 다행스러운 사태라고 볼 수 있을 것 같다. 왜냐하면 비록 통상적 正規化模型 (2.14)가 그 자체로서 바람직하지 못한 문제점들을 가지고 있음에도 불구하고 그 模型에 능선회귀법을 적용해서 얻은 회귀계수추정량은, 그런 문제점들을 갖지 않는 새로운 正規化模型 (4.10)에 능선회귀법을 적용해서 얻는 회귀계수추정량중에서 적어도 $b_2^*(\alpha)$와는 일치하기 때문이다. 그러나 모형 (2.14)에서는 $b_1^*(\alpha)$를 계산할 수 없다는 문제점은 그대로 남는다.

VII. 中心調整과 規模調整

식 (2.7)에서 우리는 X_2의 正規化된 行列 X_2^*를 다음과 같이 정의하였다. 즉.

(7.1) $X_2^*=AX_2S_2^{*-1}$

여기서 X_2의 앞에 A를 곱한 것은 X_2를 中心調整(recentering)한 것이다. 즉 A를 앞곱함으로써, X_2 속의 $k-1$개의 변수들의 평균이 모두 0이 되도록 조정해 준 것이다. 이 中心調整으로 말미암아 X_2^*의 모든 列은

(7.2) $1'X_j^*=0;\ (j=2,3,\cdots,n)$

라는 성질을 가지게 된다.

또 식 (7.1)에서 AX_2의 뒤에 S_2^{*-1}를 곱한 것은 AX_2를 規模調整(rescaling)한 것이다. 즉 S_2^{*-1}를 뒤곱함으로써, AX_2 내의 $k-1$개의 列벡터들의 길이를 모두 1이 되도록 조정해 준 것이다. 이 規模調整으로 말미암아 X_2^*의 모든 列은

(7.3) $x_j^{*\prime}x_j^*=1,\ (j=2,3,\cdots,n)$

이라는 성질을 가지게 된다. X_2^*가 正規化된 行列이라는 의미는 바로 性質 (7.2)와 (7.3)을 갖는다는 의미이고, 이로부터 $X_2^{*\prime}X_2^*$가 相關係數行列이라는 성질이 유도되는 것이다.

그러나 x_1 즉 1의 正規化된 벡터 x_1^*를 식 (4.8)에 의하여

(7.4) $x_1^*=\frac{1}{\sqrt{n}}1$

로 정의할 때, 그 正規化의 의미는 X_2^*의 正規化의 의미와는 다르다. 즉 x_1^*는 中心調整이 생략되고, 다만 規模調整이 이루어져서

(7.5) $x_1^{*\prime}x_1^*=1$

이라는 성질을 가질 뿐이다. 그리고 우리가 中心調整을 해 주지 않은 이유는, 식 (2.11)에서 본 바와 같이, 그렇게 할 때, x_1(즉 1)은 0벡터로 되고 말기 때문이었다.

한편 正規化를 위하여 모든 變數에 대해서 중심조정과 규모조정을 동시에 해주어야 할 것인가 아닌가의 문제를 제기해 보면, 제 2 절의 논의에서 본 바와 같이 상수항 변수 1을 포함하여 모든 변수를 中心調整해 줄 때, 식 (2.14)와 같은 바람직하지 않은 正規化模型을 얻을 수 밖에 없었다는 사실에 想到하게 된다. 그리고 또 그런 바람직하지 않은 결과를 얻게 된 것은 순전히 중심조정을 위하여 行列 A를 앞곱한 때문임을 알 수 있다. 즉 중심조정을 위하여 모형 (2.1)의 "앞에" A를 곱함으로써, 교란항이 $A\varepsilon$로 되어, 그 分散行列이 σ^2A

로 되었고, 또 상수항변수 벡터 1의 "앞에" A가 곱해짐으로써, 상수항이 탈락하게 되었던 것이다. 그러므로 中心調整이 이루어지더라도 모형 (2.1)의 "앞에" A가 곱해지는 방식이 아니라면, 모형 (2.14)의 바람직하지 못한 두가지 문제점이 동시에 사라지게 된다. 우리의 새로운 正規化模型 (4.10)은 바로 그렇게 하여 문제점을 해결한 것이다. 즉 y와 ε에 영향을 미치지 않으면서 다음과 같은 과정, 즉

(7.6) $X\beta=X(S^{*-1}S^{*})\beta=(XS^{*-1})(S^{*}\beta)=X^{*}\beta^{*}$

에 의해서 正規化의 목적을 달성한 것이다. 새로운 模型의 뛰어난 점은 S^{*-1}를 X에 "뒤곱"함으로써, 바람직하지 못한 문제점을 제기함이 없이, X_2에 대한 중심조정과 규모조정을 동시에 실현하였다는 점일 것이다.

VIII. 常數項이 없는 模型의 正規化

표준선형모형 (2.1) 즉,

(8.1) $y=X\beta+\varepsilon,\ \varepsilon\sim(0,\ \sigma^2I)$

에서 우리는 지금까지 X의 첫째 列이 상수항 변수 벡터 1이라고 가정하였다. 여기서는 X의 k개의 변수 중 어느 것도 상수항 변수가 아닌 모형, 즉 상수항이 없는 모형을 다루기로 한다.

既存의 능선회귀에 관한 文獻에서는 이 경우에 관한 논의가 없었다고 보아야 할 것 같다. 왜냐하면 통상적 正規化模型 (2.14)는 상수항을 포함하고 있지 않지만 사실은 이 모형이 상수항을 가지는 모형 (2.1)로부터 유도된 것이었다. 그리고 모형 (2.14)는 모형 (2.1)에서 상수항을 제외시킨 가상적 모형

(8.2) $y=X_2\beta_2+\varepsilon,\ \varepsilon\sim(0,\ \sigma^2I)$

의 앞에 A를 곱하고, X_2와 β_2 사이에 $S^{*}{}_2^{-1}S_2^{*}$를 곱하여 넣는 방법으로도 얻어질 수 있다. 그러나 그것은 제 4 절의 논의에서 알 수 있듯이, 그러한 가상적인 모형이 아니라 모형(2.1)과 관계를 맺는다고 보는 것이 더욱 합리적이다. 즉 상수항이 들어 있는 모형으로부터 유도 되었다고 보아야만 문제들이 무리없이 해결된다. 이와 관련하여 우리가 상기해 보아야 할 것은 OLS의 경우에도 中心調整 후에 不變性을 가지는 모형은 상수항을 포함하는 모형이라는 점일 것이다. 그러므로 正規化모형 (2.14)가 상수항을 포함하지 않는다는 사실은, 능선회귀에서 상수항이 없는 모형을 다룬다는 것이 아니라, 반대로 반드시 상수항을 포함하는 모형만을 다룬다고 보는 것이 옳다고 말할 수 있는 것이다.

이상의 논의에서 분명한 것은 모형 (8.1)에 대해서 기계적으로 中心調整과 規模調整을

해 주면, 그 결과는 상수항 하나를 더 추가한 모형의 正規化와 같은 결과가 된다는 사실일 것이다. 또 모형 (8.1)에는 상수항이 없기 때문에 새로운 正規化模型 (4.10)의 형태로 변형하는 것도 불가능하다. 그러면 常數項이 없는 모형의 正規化는 어떻게 정의하는 것이 合理的일 것인가? 그것은 變數의 規模調整만으로 正規化하는 것이라고 생각된다. 그 論據를 설명해 보기로 하자.

行列 X의 j번째 列을 x_j라하고, 그 規模調整된 벡터 x_j^*를 다음과 같이 정의한다. 즉

(8.3) $x_j^{**}=x_j/s_j^{**}$

단 s_j^{**}는 다음과 같이 정의된다. 즉,

(8.4) $s_j^{**}=\sqrt{x_j'x_j}=(\sum_{i=1}^{n}x_{ij}^2)^{\frac{1}{2}}$

그러면 x_j^{**}는 길이가 1인 벡터로 된다. 즉,

(8.5) $x_j^{**\prime}x_j^{**}=1,\ j=1,2,\cdots,k$

여기서 行列 S^{**}를 다음과 같이 정의하자.

(8.6) $$S^{**}=\begin{bmatrix} s_1^{**} & & & 0 \\ & s_2^{**} & & \\ & & \ddots & \\ 0 & & & s_k^{**} \end{bmatrix}$$

그러면 行列 X의 規模調整된 行列 X^{**}는 다음과 같이 정의된다. 즉,

(8.7) $X^{**}=XS^{**-1}=[x_1^{**},\ x_2^{**},\cdots,x_k^{**}]$

우리는 여기서 모형 (7.1)의 正規化모형을 行列 X^{**}를 써서 다음과 같이 정의하고자 한다. 즉,

(8.8) $y=X^{**}\beta^{**}+\varepsilon,\ \varepsilon\sim(0,\ \sigma^2I)$

단 β^{**}는

(8.9) $\beta^{**}=S^{**}\beta$

로 정의된다. 모형 (7.8)은

(8.10) $X\beta=XS^{**-1}S^{**}\beta=X^{**}\beta^{**}$

의 관계에 의해서 식 (8.1)로부터 유도되므로 모형 (4.10)이 그러했던 것처럼 아무런 바람직하지 못한 문제점도 발생하지 않는다. 다만 行列 $X^{**\prime}X^{**}$가 어떤 의미에서 正規化의 조건을 충족할 수 있는가를 확인하면 된다.

行列 $X_2^{*\prime}X_2^*$는, 대각원소가 모두 1이고, (i,j)번째 원소인 $x_i^{*\prime}x_j^*$는 i번째 변수 x_i와 j번째 변수 x_j간의 相關係數로 해석할 수 있기 때문에, 그 行列을 相關係數行列이라고 부를

수 있었다. 또 $X^{*\prime}X^{*}$ 역시 대각원소는 모두 1이고 비대각원소는 첫째 행과 첫째열의 경우 약간의 의미의 수정을 가하면 역시 상관계수행렬이라고 부를 수 있었다. 그러면 행렬 $X^{**\prime}X^{**}$는 어떠한가. 이를 전개해 보면 다음과 같다.

$$(8.11)\quad X^{**\prime}X^{**}=\begin{bmatrix} 1 & x_1^{**\prime}x_2^{**} & x_1^{**\prime}x_3^{**}\cdots x_1^{**\prime}x_k^{**} \\ x_2^{**\prime}x_1^{**} & 1 & x_2^{**\prime}x_3^{**}\cdots x_2^{**\prime}x_k^{**} \\ \cdots & \cdots & \cdots \\ x_k^{**\prime}x_1^{**} & x_k^{**\prime}x_2^{**} & x_k^{**\prime}x_3^{**}\cdots \quad 1 \end{bmatrix}$$

즉 이 행렬 역시 대각원소는 모두 1이다. 그러면 비대각원소 $x_i^{**\prime}x_j^{**}$들은 상관계수라고 해석할 수 있겠는가?

行列 $X_2^{*\prime}X_2^{*}$의 원소 $x_i^{*\prime}x_j^{*}$를 相關係數로 볼 수 있는 것은, x_i^{*} 및 x_j^{*}가 모두 中心調整과 規模調整이 된 벡터들이기 때문이다. 그리고 슈바르쯔 不等式

$$(8.12)\quad (x_i^{*\prime}x_i^{*})\cdot(x_j^{*\prime}x_j^{*})\geq(x_i^{*\prime}x_j^{*})^2$$

으로부터

$$(8.13)\quad -1\leq x_i^{*\prime}x_j^{*}\leq 1$$

라는 부등식을 얻는다. 그런데 슈바르쯔不等式 (8.12)는 x_i^{*} 및 x_j^{*}가 規模調整된 벡터이기 때문에 (8.13)으로 변형되는 것이지, 이들이 中心調整된 벡터라는 사실과는 無關하다. 따라서 우리는 行列 X^{**}의 規模調整된 벡터 x_i^{**}와 x_j^{**}에 관해서도

$$(8.14)\quad -1\leq x_i^{**\prime}x_j^{**}\leq 1$$

라는 부등식이 성립하는 것을 쉽게 확인할 수 있다. 따라서 中心調整되지 않은 벡터들로 구성되는 $x_i^{**\prime}x_j^{**}$를 相關係數라고 부를 수는 없지만 (8.14)의 부등식이 성립하는 사실을 고려하여, 이를 準相關係數(pseudo-correlation coefficient)라고 부를 수는 있을 것이다. 이 개념에 관한 기존문헌의 用語는 잘 알 수 없으나, 計量經濟學에서 中心調整이 이루어지지 않은 상태에서의 決定係數 (R^2)가 이와 극히 유사한 개념인데 이를 타일[1971, p. 164, 178]과 아메미야[1985, p. 6]는 통상적인 決定係數와 구별없이 그대로 "決定係數" 또는 R^2라는 말을 사용하고 있다. 이것을 우리의 用語에 따라 바꿔 부른다면 "準決定係數"라고 부를 수 있을 것이다.

요컨대 正規化模型 (8.8)은 $X^{**\prime}X^{**}$가 準相關係數行列이라는 의미에서 正規化된 모형이다. 이를 위해서 X속의 모든 變數는 規模調整되었다. 한편 正規化模型 (4.10)은 常數項變數는 規模調整하고, 기타 설명변수들은 中心調整 및 規模調整을 함으로써 모형을 正規化하였다. 모형 (4.10)은 상수항이 포함된 모형에만 적용될 수 있다. 모형 (8.8)은 상수항이 포함되지 않은 모형에 적용될 수 있다. 그러나 이상의 설명을 음미해 보면, 상수항이 들어

있는 경우에도 모형 (8.8)을 적용할 수 있다. 그러나 모형(4.10)을 적용하는 경우와 다른 결과를 얻는다. 상수항이 들어 있는 경우 두 모형 중 어느 것을 사용할 것이냐는 分析者의 취향에 따르는 것이라고 말할 수 있을 것이다. 그러나 상수항이 들어 있는 모형에서는 결정계수를 중심조정된 상태에서 계산하는 것이 보통이듯이, 상수항이 들어 있는 모형에서는 능선회귀도 모형 (4.10)을 쓰는 것이 보다 일반적인 관행으로 되어야 할 것이다.

參 考 文 獻

Amemiya, T., *Advanced Econometrics*, Harvard University Press, 1985.

Hoerl, A.E., and R.W. Kennard, "Ridge Regression: Biased Estimation of Nonorthogonal Problems," *Technometrics* 12, Feb. 1970.

朴聖炫, 『回歸分析』, 大英社(서울), 1981.

Sawa, T., 佐和隆光, 『回歸分析』, 朝倉書店(東京), 1979.

Theil, H., *Principles of Econometrics*, Wiley, 1971.

Vinod, H.D., "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares," *Review of Economics and Statistics*, Feb. 1978.