回歸分析에 있어서의 多共線性과 名稱을 保全시키는 資料變換 技法

兪 浣*

I. 序 論

y를 從屬變數라 하고 x를 ρ次元의 獨立變數, β를 m次元의 回歸係數(regression coefficients)라 하며 ε을 誤差項(error term)이라고 하면 回歸方程式은 아래와 같이 表示된다.[1]

$$y_t=\beta' x+\varepsilon_t \tag{1}$$

n個의 觀察資料를 가지고 흔히 β의 推定量은 最尤法(maximum likelihood method)에 의하여 다음과 같이 구하여진다.

$$b=(X'X)^{-1}X'y \tag{2}$$

이와 같이 b를 구하는 과정에서 多共線性(mullicoltinearity) 問題가 흔히 發生하게 되며, 이를 피하는 方法으로서 要因分析(factor analysis), 主成分分析(principal component analysis), 또는 段階的回歸分析(stepwise regression analysis)등이 흔히 이용된다. 위와 같은 방법은 여러가지 長點이 있으나, 만일 獨立變數의 이름이 重視되며 또 그 중요성에 있어 優劣을 가리기 어려운 경우에 곤란을 겪게 되는 수가 있다.

예컨데 y가 個人의 交通量이라 하고 x_1을 交通時間, x_2를 交通費라 하고 回歸線에 의한 交通量 推定公式을 推定하여 交通時間과 交通費의 代替效果(substitution effect)를 交通時間의 價値(VOT; value of time)로 算定한다면, 이는 x_1과 x_2의 對應된 係數(coefficients) b_1과 b_2의 比로써 나타나게 된다.

$$\begin{aligned} VOT&=-\frac{\partial x_2}{\partial x_1} \\ &=\frac{\partial y/\partial x_1}{\partial y/\partial x_2} \\ &=\frac{b_1}{b_2} \end{aligned} \tag{3}$$

위와 같은 연구과정에서 교통의 起點 및 終點이 다른 觀測資料를 사용했을 경우[2] 흔히 交

*延世大學校 工科大學 助敎授

1) 보통 回歸分析(regression analysis)에서 나오는 常數는 편의상 무시하였다.

2) 위와 같은 接近方法의 타당성은 여기서 考慮되고 있지 않고 모든 변수는 陽函數로 간주한다.

通時間 및 交通費가 모두 交通距離에 직접 관계되어 多共線性(multicollinearity)의 問題를 惹起시키나, 연구목적이 VOT를 求하는데 있다면 交通時間과 交通費라는 獨立變數중에 하나를 제거할 수 없으며 두 變數의 重要性의 優劣도 가릴 수 없게 된다.

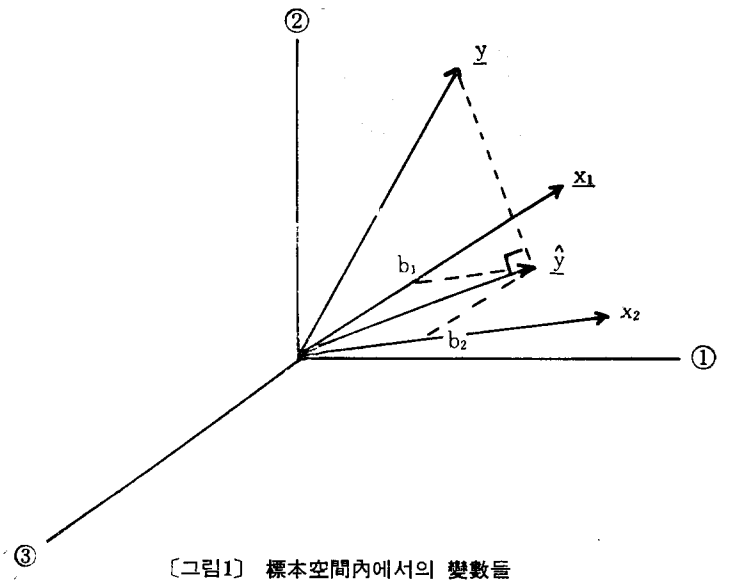
이 論文은 위에서 例示한 바와 같은 多共線性(multicollinearity) 問題를 취급하기 위한 代案으로서, 稜形回歸技法(ridge regression technique)과 要因分析技法(factor analytic technique)을 소개하는데 主目的을 두었다.

II. 回歸線과 多共線性 問題

널리 쓰이고 있는 方法으로, 回歸線을 標本空間內에서의 透射(projection)하는 것으로 說明하면 다음과 같다. 표본규모가 셋인 표본공간에서 다음과 같은 자료수집이 이루어졌다고 하자.

$$\begin{pmatrix} y_1 & x_{11} & x_{12} \\ y_2 & x_{21} & x_{22} \\ y_3 & x_{31} & x_{32} \end{pmatrix} = (\boldsymbol{y}\ \boldsymbol{x}_1\ \boldsymbol{x}_2) \tag{4}$$

위의 자료를 標本空間內에서 〔그림 1〕과 같이 表示할 수 있다.



〔그림1〕 標本空間內에서의 變數들

〔그림 1〕에서 표시된 $\boldsymbol{y}$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$는 각각 벡타(vector)로 표시되고 回歸線에서 얻어지는 $\boldsymbol{y}$의 推定量 $\hat{\boldsymbol{y}}$은 $\boldsymbol{y}$를 $\boldsymbol{x}_1$과 $\boldsymbol{x}_2$가 만드는 空間에서 直角의 透射(projection)로 표시되며 $\boldsymbol{x}_1$과 $\boldsymbol{x}_2$의 凸型結合(convex combination)으로 표시된다.

$$\hat{y}=b_1x_1+b_2x_2 \tag{5}$$

이때 x_1과 x_2가 직각보다 좁은 角을 이루는 벡타로 나타날 때 x_1과 x_2가 만드는 空間이 좁아짐으로써 多共線性問題가 발생하며 b_1과 b_2의 값이 안정되지 못하게 된다. 그리하여 多共線性問題를 處理하는 方法으로서 x_1과 x_2를 體系的으로 서로 벌려주는 方法을 생각 할 수 있다. 그렇게 함으로써 b_1과 b_2의 값을 安定시켜 주게 된다.

III. 稜形回歸技法(ridge regression technique)

호엘과 케나드(1970 a)에 의해 體系化된 稜形回歸(ridge regression)方法은 平均平方誤差를 最小化(minimizing mean square error)하는 方法으로 설명되며 비노드(1977)에 의해 아래와 같이 표시 되었다. (6)

$$\begin{aligned} MSE(\boldsymbol{b}) &= E[(\boldsymbol{b}-\boldsymbol{\beta})^2] \\ &= E[\{\boldsymbol{b}-E(\boldsymbol{b})\}^2]+E[\{E(\boldsymbol{b})-\boldsymbol{\beta}\}^2] \\ &= \text{trace } Var(\boldsymbol{b})+[bias(\boldsymbol{b})]^2 \end{aligned}$$

稜形回歸推定量 $\boldsymbol{b}^r$은 가능한 모든 推定量 $\boldsymbol{B}$들 중에서 分散을 極小化하는 방법으로 찾게 된다.

$$\min \boldsymbol{B}'\boldsymbol{B} \tag{7}$$

$$\text{s.t. } (\boldsymbol{B}-\boldsymbol{b})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{b})=0$$

위에서 보는 바와 같이 $\boldsymbol{b}^r$은 약간의 偏倚(bias)를 허용함으로써 分散을 많이 줄이는 方法으로 위의 極小化問題의 答인 稜形回歸推定値는 아래와 같이 포시 된다.

$$\boldsymbol{b}^r=(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{8}$$

$$k\geq 0$$

따라서 $\boldsymbol{b}^r$은 偏倚와 分散을 交換하는 方法으로 설명된다.

위와 같은 방법은 자료처리의 과정에서 그 설명의 補完이 必要하다. 다음은 이러한 稜形回歸推定量을 자료의 變換 후에 얻는 OLS 推定量으로 설명, 근본자료에서 어떠한 變換을 가져 오는 것인가를 살펴 보기로 한다.

稜形回歸推定量은 원래 자료 $\boldsymbol{X}$를 稜形回歸를 前提條件으로, 새로이 變換된 자료 $\boldsymbol{Z}$와의 유클리디안(Euclidean)距離를 極小化함으로써 얻어지는 OLS 推定量으로 解析할 수 있다.

$$\min (\boldsymbol{X}-\boldsymbol{Z})'(\boldsymbol{X}-\boldsymbol{Z}) \tag{9}$$

$$\text{s.t. } (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'=(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$$

위의 極小化問題의 制約式(constraint)에 $(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})$를 우선 곱해 주고 뒤에 $\boldsymbol{Z}$를 곱하면 다음과 같은 문제로 변하게 된다.

$$\min(\boldsymbol{X}-\boldsymbol{Z})'(\boldsymbol{X}-\boldsymbol{Z}) \tag{10}$$

s.t. $X'Z=X'X+kI$

이상의 문제를 라그랑쥬 行列(Lagrangian matrix) λ를 사용하여 풀면 다음과 같은 결과를 얻는다.

$$Z=X+\frac{1}{2}X\lambda \tag{11}$$

$$\lambda=2k(X'X)^{-1}$$

따라서 Z는 다음과 같이 된다.

$$Z=X+kX(X'X)^{-1} \tag{12}$$

위의 Z는 極小問題 (10)과 (11)의 答이 된다. $(X'X+kI)^{-1}X'$이 $(Z'Z)^{-1}Z'$로 됨을 證明하기 위하여 $X'X$의 스펙트랄 分解(spectral decomposition)를 $V\Delta V'$로 표시하자.

$$X'X=V\Delta V' \tag{13}$$

위에서 Δ는 固有値(eigenvalue)의 對角行列(diagonal matrix)를 나타내며 V는 Δ에 相關된 固有벡타(eigenvector)의 行列이 되며 $(X'X+kI)^{-1}X'$는 아래와 같이 표시된다.

$$\begin{aligned}(X'X+kI)^{-1}X' &= V(\Delta+kI)^{-1}V'X' \\ &= V(\Delta+kI)^{-1}(I+k\Delta^{-1})^{-1}(I+k\Delta^{-1})V'X' \\ &= V[(\Delta+kI)(I+k\Delta^{-1})]^{-1}V'V(I+k\Delta^{-1})V'X' \\ &= (V\Delta V'+2kI+k^2V\Delta^{-1}V')^{-1}(I+kV\Delta^{-1}V')X' \\ &= [X'X+2kI+k^2(X'X)^{-1}]^{-1}(X'+k(X'X)^{-1}X') \\ &= [(X+kX(X'X)^{-1})'(X+kX(X'X)^{-1})]^{-1}(X+kX(X'X)^{-1}) \\ &= (Z'Z)^{-1}Z'\end{aligned} \tag{14}$$

따라서 稜形推定量은 아래와 같이 OLS 推定量으로 표시될 수 있다.

$$b^r=(Z'Z)^{-1}Z'y \tag{15}$$

따라서 稜形推定量은 원래의 관측자료 X를 公式(12)에 의해 Z로 變換함으로써 x_1과 x_2가 이루는 角을 z_1과 z_2로 分散하여 이에 해당되는 安定된 最尤推定量(maximum likelihood estimator)을 얻는 방법으로 요약할 수 있다. 이때 공식(12)는 X를 Z로 變換하는 變換公式이 된다.

IV. 要因分析技法(factor analytic technique)

稜形推定量(ridge estimator)은 원래의 觀測資料를 變換하여 얻은 OLS 推定量으로 간주할 수 있다는 장점에 반하여 원래의 자료의 位置가 변한다는 약점을 가지고 있다. 이러한 약점을 감안, 아래와 같은 要因分析技法을 發展시켰다. 아래의 방법은 單一要因模型(one factor model)으로 一般要因分析技法을 逆利用한 방법이다.

만일 서로 相關關係에 있는 두 개의 變數 x, z를 서로 獨立된 要素x_I, z_I와 共同要素

$x_c=z_c=c$로 표시할 수 있다고 가정하자

$$\begin{aligned} x&=x_I+c \\ z&=z_I+c \\ \mathrm{Cov}(x_I,z_I)&=0 \end{aligned} \tag{16}$$

公式(1)에서 나타난 回歸線의 공식은 아래와 같이 표시할 수 있다.

$$y=\beta_x x_I+\beta_z z_I+(\beta_x+\beta_z)c+\varepsilon \tag{17}$$

여기서 새로운 변수 w를 아래와 같이 定義하여 보자.

$$w=(\beta_x+\beta_z)c \tag{18}$$

公式(18)을 (17)에 代入하면 다음과 같은 式을 얻는다.

$$y-w=\beta_x x_I+\beta_z z_I+\varepsilon \tag{19}$$

두 개의 獨立變數 x와 z를 V로 표시할 때, 즉

$$\begin{aligned} V&=(x\ z) \\ V_I&=(x_I\ z_I) \\ V_c&=(c\ c) \end{aligned} \tag{20}$$

와 같이 표시된다고 하면, 公式(19)에 해당하는 OLS 推定量 $b^f=(\hat{\beta}_x\ \hat{\beta}_z)$는 아래와 같다.

$$b^f=(V_I'V_I)^{-1}V_I'(y-w) \tag{21}$$

위의 公式은 $\mathrm{Cov}(x_I,z_I)=0$라는 條件을 가진 單一要因模型(one factor model)이다. 위에서는 殘差(residual)인 $(y-w)$를 사용함으로써 x_I와 z_I가 본연의 變數 이름을 그대로 유지시키는 장점을 갖고 있다. 그리하여 說明變數 x_I와 z_I가 可能한 限, 많은 부분을 설명할 수 있게 하기 위하여는 共同要素인 c를 可能한 限 적게 해 주어야 한다. 즉

$$\begin{aligned} &\min\ V_c'V_c=(V-V_I)'(V-V_I) \\ &\text{s.t. } \mathrm{Cov}(x_I,z_I)=0 \\ &\qquad V=V_I+V_c \end{aligned} \tag{22}$$

위의 二次型(quadratic) 問題를 풀면 아래와 같은 答을 얻는다.

$$V_I=V(I+\Delta)^{-1} \tag{23}$$

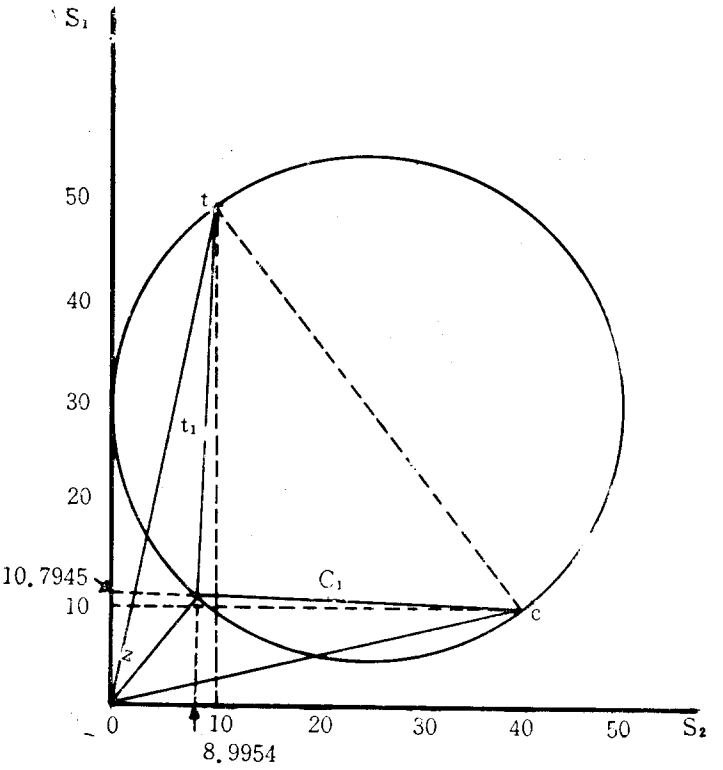
여기서 Δ는 라그랑쥬(Lagrangian) 行列로 다음 式을 만족시킨다.

$$V'V=(I+\Delta)'(I+\Delta) \tag{24}$$

위의 式은 標本空間(sample space)에서 쉽게 설명 될 수 있다. 2개의 標本이 아래와 같이 주어 졌다고 假定하고 그림 2를 참조하면,

$$\begin{aligned} x&=(50\ 10)' \\ z&=(10\ 40)' \end{aligned} \tag{25}$$

여기서 x_I와 z_I가 서로 獨立되었다면 標本空間에서 直交(orthogonal)로 나타나며 x와 z를 통과하는 無限次元圓(infinite dimensional circle)의 표면에 있는 모든 點들은 이를 만족시킨다



〔그림 2〕 標本空間內에서 나타낸 가상자료의 要因分析的(factor analytic) 전환결과

(제약식이 주어지는 조건). 이러한 모든 點들 중에서 共同要素 c가 最小로 되는 ——x_1와 z_1가 最大로 되는—— 點은 原點에서 最短距離에 있는 點을 말한다. 그러므로

$$c=(10.7945\ \ 8.9954)' \tag{26}$$

로 되고,

$$V_1=(x_1\ z_1) \tag{27}$$
$$=\begin{pmatrix}39.2055 & -0.7945\\ 1.0046 & 31.0046\end{pmatrix}$$

과 같이 된다. 참고로 원래의 共分散行列(variance-covariance matrix)과 獨立要素의 共分散行列을 살펴 보면 다음과 같다.

$$\mathrm{Cov}(V)=\begin{pmatrix}2600 & 900\\ 900 & 1700\end{pmatrix} \tag{28}$$
$$\mathrm{Cov}(V_1)=\begin{pmatrix}1538 & 0\\ 0 & 962\end{pmatrix}$$

V. 結 論

두 개의 變數의 代替效果(substitution effect)를 硏究하기 위하여 需要 또는 供給의 模型을 만들었을 경우 이에 關聯된 變數들의 이름이 중요시 된다. 실제 觀測資料를 사용하였을 경우 흔히 일어나는 多共線性(multicollinearity) 問題를 다루기 위한 代案으로써 線型回歸線을 例로 들어 稜形回歸技法(ridge regression technique)과 要因分析技法(factor analytic technique)을 소개하였으며 이에서 얻어 지는 係數(coefficient)를 OLS 推定値로 설명하기 위하여 원래의 자료를 變換하였다. 실지 需要와 供給의 模型이 非線型일 경우 一般的으로 稜形回歸나 要因分析을 쓰지 못한다는 점을 감안, 이러한 방법을 자료의 變換方法으로 설명함으로써 非線型模型에서도 多共線性問題를 위하여 稜形回歸技法(ridge regression technique)이나 要因分析技法(factor analytic technique)을 사용할 수 있도록 하였다.

參 考 文 獻

[1] Bartels, C., *Operational Statistical Methods for Analyzing Spatial Data*, Mimeo, University of Groningen, The Netherlands, 1977.

[2] Bellman, R., *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1970.

[3] Cohen, J. et al, *Applied Multiple Regression Correlation for Behavioral Science*, Lawrence Erlbaun Associates, Hillside, New Jersey, 1975.

[4] Dempster, A.P., "Alternative to Least Squares in Multiple Regression," in Kale, D.G. and Gupta, R.P. (eds), *Multivariate Statistical Inference*, North-Holland Publishing Company, Amsterdam, The Netherlands, (1973), 27-40.

[5] Dempster, A.P. et al. "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association*, Vol. 72(1977), 77-91.

[6] Efron, B. and Morris, C., "Data Analysis Using Stein's Estimator and Its Generalization," *Journal of the American Statistical Association*, Vol. 70(1975), 311-319.

[7] Hoerl, A.E. and Kennard, R.W., "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics* Vol. 12, No. 1(1970a), 55-67.

[8] Hoerl, A.E. and Kennard, R.W., "Ridge Regression Applications to Nonorthogonal Problems," *Technometrics*, Vol. 12, No. 1(1970b), 69-82.

[9] James, W. and Stein, C., "Estimation with Quadratic Loss," in Neyman, J. (ed.), *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and*

Probability 1, University of California Press, Berkeley, 1961.

[10] Lindley, D.V. and Smith, A.F.M., "Bayes Estimates for the Linear Model," *Journal of Royal Statistical Society*, Vol. B34, No. 1(1972), 1-4.

[11] Marquardt, D.W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, Vol. 12, No. 3(1970), 59-61.

[12] Smith, T.E., "A Data Perturbation Approach to Generalized Ridge Regression," (forthcoming) (1979)

[13] Vinod, H.D., *A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Squares*, Part I and Part II, Mimeo, (1977)

[14] Yu, W., *Contributions to Applications for Linear Logit Models in Transportation Research*, Ph.D. Thesis, University of Pennsylvania, Philadelphia, (1978)