韓國經營科學會誌
第28卷 第3號
2003年 9月

상관관계가 강한 독립변수들을 포함한 데이터 시스템 분석을 위한 편차 - 복구 알고리듬*

이 미 영**

Biased-Recovering Algorithm to Solve a Highly Correlated Data System*

Miyoung Lee**

■ Abstract ■

In many multiple regression analyses, the "multi-collinearity" problem arises since some independent variables are highly correlated with each other. Practically, the Ridge regression method is often adopted to deal with the problems resulting from multi-collinearity. We propose a better alternative method using iteration to obtain an exact least squares estimator. We prove the solvability of the proposed algorithm mathematically and then compare our method with the traditional one.

Keyword : Ridge Regression, Least Squares, Eigenvalue, Correlation Matrix, Iteration

1. 서 론

기존의 기업에서의 전략이나 산업현장에서의 공정관리 및 문제해결은 오랜 경험을 가진 전문가의 판단에 의해 이루어졌다. 그러나 현대사회는 기업뿐 아니라 산업전반에 걸쳐서 경쟁이 심화되고, 경쟁력의 바탕이 정보 전략에 있다는 점이 인식됨에 따라 정보의 중요성에 대한 인식이 심화되고 있다. 이에 따라 기존에는 그저 기록용 수치로만 여겨졌던 데이터가 의미를 가지게 되었다. 이제 더 많은

논문접수일 : 2003년 6월 30일 논문게재확정일 : 2003년 7월 30일
* 이 논문은 '2002년도 건국대학교 신임교원연구비 지원에 의한 논문임.
** 건국대학교 경영대학 경영정보학과 조교수

데이터가 수집되고 그 데이터로부터 가치 있는 정보를 찾아내고자 하는 노력이 이루어지고 있다. 이러한 사회의 추세에 따라 데이터의 중요성이 인식되고 있고, 이러한 데이터를 수집하는 일이 주요한 기업의 업무 중의 하나가 되고 있으며, 이렇게 축적된 데이터로부터 정보를 추출하기 위한 여러 데이터 분석에 관한 연구가 되어지고 있다. 이러한 연구의 동향을 크게 두 가지로 나눈다면 새로운 분석 방법론의 제시에 관한 연구 및 전통적인 분석방법을 이용한 적용사례를 다루는 연구로 나눌 수 있다.

데이터 수집, 정제, 분석활동 전반을 일컫는, 소위 데이터 마이닝 이라고 불리는 이러한 방법 활동에 대한 정의를 인용해 보면 "대량의 데이터로부터 새롭고 의미 있는 정보를 추출하고 의사결정에 활용하는 작업"이라고 한다[1]. 데이터로부터 의미있는 정보를 추출하기 위해 우선 되어야 하는 작업은 데이터 선택이며 그 다음으로 데이터의 정제 및 보완이라고 할 수 있다. 그 다음으로 이어지는 작업이 데이터를 분석할 수 있는 모습으로 변환을 하는 것이다. 이러한 준비된 데이터를 이용하여 여러 가지 기법을 이용하여 분석하는 작업이 이루어진다. 방법론으로서 의사결정나무(Decision Tree), 신경망(Neural Networks), 동시발생 매트릭스(Co-Occurrence Matrix), 군집화(Clustering) 등의 많은 데이터 분석 기법이 사용되고 있으나 가장 일반적으로 쓰이고 있는 방법 중의 하나가 회귀분석 방법이다. 회귀분석을 하기 위해 요구되는 기본적인 가정들이 있는데, 예를 들면 그 데이터들이 정규분포를 따라야 하고, 독립변수의 독립성이 보장되어야 한다. 만약 이러한 가정이 잘 성립되지 않으면 회귀분석의 결과를 신뢰할 수 없는 상황이 되거나 효율적인 결과 도출이 어렵다. 그러나 기업이나 산업현장에서 독립변수로 다뤄져야 하는 데이터들은 이러한 가정이 성립되지 않는 경우가 대부분이다. 본 논문에서는 이러한 회귀분석의 문제점, 즉, 가정이 성립되지 않는 데이터들을 이용하여 분석함으로써 생기는 오류를 최소화하고 그러한 데이터로부터 신빙성 있는 답을 도출해 낼 수 있는, 보완된 알고리듬을 제시하고자 하는 것이 목적이다. 이 논문의 2장에서는 일반적으로 알려진 회귀분석 방법론에 대하여 이야기하고 3장에서는 이러한 방법론의 단점을 보완하기 위한 새로운 방법론을 제시하고 이 새로운 방법론을 체계적으로 데이터 시스템에 적용할 수 있는 알고리듬을 제시한다. 또한 제시된 알고리듬의 적용가능성 및 신빙성을 수학적으로 증명한다. 4장에서는 제시된 방법론을 데이터에 적용하여 그 결과를 다른 방법론과 비교 분석할 것이다.

2. 분석 방법론

회귀분석방법은 주어진 데이터 군으로부터 우리가 원하는 결과에 미치는 영향을 계수 값을 통해 각 데이터별 수치함수로 가장 잘 보여주는 방법이라 할 수 있다. 그래서 이 회귀분석방법을 잘 적용하기 위한 여러 가지 방법론이 제기 되고 있다[5, 7].

이 회귀분석방법의 기초가 되는 것이 결국은 행렬문제를 푸는 것으로 귀착되며, 사용되고 있는 대부분의 방법론의 기저에는 통계 툴을 바탕으로 하는 행렬문제로 귀착되고 있다. 일반적으로 행렬문제는 간단히 역 행렬을 구하면 되는 것으로 인식되고 있으나, 이것이 항상 성립하는 것은 아니다. 행렬의 성질에 따라서는 전혀 풀 수 없거나 우리가 원하는 정보가 아닌, 왜곡된 값을 구해주는 경우도 많다.

주어진 데이터로부터 관계식을 도출하는 방법으로서 가장 일반적으로 쓰이는 방법이 최소제곱 추정 값에 의해 관계식을 도출해 내는 방법이다.

다음은, i 번째 데이터를 이용한 관계식을 나타낸다.

$$y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_n x_{in} + \varepsilon_i$$
$$i = 1, 2, \cdots, k \qquad k : \text{데이터 개수}$$

이때, 위의 관계식은 표준화 되어있는 것으로 간주한다. 즉,

$X_j = (x_{1j}, x_{2j}, \cdots, x_{kj})'$라고 할 때, $x_{ij} \sim (x_{ij}$

$-E(X_j)/(Var(X_j))^{1/2}$를 가정한다. 앞에서의 관계식을 행렬을 이용하여 다시 표현하면,

$$Y = X\beta + \varepsilon \tag{1}$$

이라고 할 수 있다.

이때, Y는 $k\times 1$, X는 $k\times n$ 그리고 β는 $n\times 1$ 행렬이다. ε은 측정오차를 나타내며 $k\times 1$ 행렬이 된다.

이제, 주어진 데이터로부터 독립변수와 종속변수의 관계를 밝힌다는 것은 $Y = X\beta + \varepsilon$에 최소제곱 추정 방법을 적용하여 $Y - X\beta$의 제곱 값을 최소로 하는 β 값을 구하는 것이다. 이것은 주어진 데이터를 가장 잘 표현하는 등식을 만드는 작업이라 할 수 있다. 이 최소제곱 추정 값을 적용하는 과정에 대해 간단히 설명해 보면 다음과 같다.

우선 식 (1)과 동치인 다음과 같은 식 (2)를 생성한다.

$Y - X\beta$의 식 에서 변화하는 벡터 값 β에 대하여 최소 값을 가져야 하므로 벡터 $Y - X\beta$의 제곱 값을 벡터 β의 각 인수에 대하여 미분하여, 그 제곱 값을 최소로 하는 β 값을 찾는 것이다[2]. 즉

$$\frac{\partial}{\partial \beta_i}|Y - X\beta|^2 = 0 \tag{2}$$

을 만족하는 β를 찾는데, 그 값이 우리가 원하는 해 $\hat{\beta}$이다.

위의 식 (2)을 풀어 다시 행렬방식으로 표현하면 결과 값 $\hat{\beta}$는 다음과 같이 나타내어진다.

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$

이때, 독립 변수 간의 상관관계가 클수록 이 최소제곱 추정 값은 신빙성이 떨어지며 몇 개의 데이터를 쓰느냐에 따라 값이 달라지기도 한다. 이것을 통계학적으로 설명하면 상관관계 행렬인 $X'X$가 단위 행렬의 형태에서 멀어질수록, 즉 상관관계가 많이 일어날수록 최소제곱 추정 값은 관측값의 변화에 매우 민감하다고 할 수 있고[3], 이것을 다시 수학적으로 표현하면 다음과 같다. 독립변수들이 서로 완전독립이 아니고 상관관계가 있다면, 관계 행렬 $X'X$의 특성 값 중에 0에 가까운 수가 나타나게 된다. 그런데 $\hat{\beta}$를 구하기 위한 행렬 방정식 (3)을 푸는데 있어서 그 해의 안정성, 즉 해에 대한 신용도 는 그 특성 값으로 정의된다. 행렬 $X'X$의 특성 값들을 다음과 같이 나열할 때

$$|\lambda_{\max}| = |\lambda_n| \geq |\lambda_{n-1}| \geq \cdots \geq |\lambda_1| = |\lambda_{\min}| > 0 \tag{4}$$

행렬방정식 (3)에서의 $\hat{\beta}$에 대한 신용도는 컨디션 넘버라고 일컫는

$$cond(X'X) = |\lambda_{\max}|/|\lambda_{\min}| \tag{5}$$

에 의해 정의된다.

이때, $|\lambda_{\min}|$이 0에 가까운 값이 된다면, 이 방법으로는, 해의 안정성과 상관없이 해가 구해지지 않는다[6].

이러한, 독립변수들 간의 비 독립성에 의해 일반적인 최소제곱 추정법을 통해 값을 구하는 것이 힘든 상황에 대한 대안으로 통계학 적으로 제안된 방법이 다음에 나오는 Ridge Regression 방법이다. 그 뒤에 따라 나오는 절에서는 Ridge Regression 방법에 대응하는 방법으로서 이 논문에서 제안하는 Biased-Recovering 알고리듬이다. 이 두 가지 방법은 모두 행렬 $X'X$의 특성 값이 음수가 아니라는 성질을 이용하고 있다.

이제 주어진 $X'X$의 특성 값들을 다음과 같은 순서로 나열하고, 앞으로 이 논문에서는 다음의 식 (6)을 전제하고 내용을 전개하기로 한다.

$$\lambda_{\max} = \lambda_n \geq \lambda_{n-1} \geq \cdots \geq \lambda_1 = \lambda_{\min} > 0 \tag{6}$$

Ridge Regression 방법은 기본적으로 우리가 풀고자 하는 행렬 방정식 (3)의 $X'X$에 약간의 변화량을 주어 $X'X$ 대신에 $X'X + kI$에 대하여 문제를 푼다[3].

이 변화된 행렬을 이용하여 구한 해는 $\widehat{\beta^*} = (X'X+kI)^{-1}X'Y$로 표현되고 $\widehat{\beta}$와 $\widehat{\beta^*}$의 관계는

$$\widehat{\beta^*} = (I+k(X'X)^{-1})^{-1}\widehat{\beta} \quad (7)$$

로 나타내어진다. $X'X$ 행렬의 문제점이 매우 작은 특성 값에 의하여, 일반적인 최소제곱법을 적용할 수 없는 것이므로 행렬 $X'X$의 특성 값이 0값에서 멀어지도록 하자는 것이다. 그러나, k 값이 너무 커질 경우 $X'X$와 $X'X+kI$는 완전히 다른 행렬이 되고, 따라서 그 두 행렬을 이용하여 구하여진 두 해의 차이는 식 (7)에서 보는 바와 같이 매우 크게 된다.

3. Biased-Recovering Algorithm

이 장에서는 0에 가까운 특성값을 가질 때에도 정확한 회귀분석결과를 도출하기 위한 알고리듬을 제시한다. 우선 데이터간에 상관도가 높은 경우의 회귀분석을 위해 Ridge regression을 이용하여 회귀분석을 하여 해을 구한다. 그러나 이 해는 위에서 언급한 바와 같이 이미 행렬에 오차를 더하여 해를 구한 것이므로 그 해 또한 오차를 포함한다. 우리는 여기서, Ridge Regression을 이용하여 구한 해를 다음에서 제시하는 알고리듬에 따라 두 세 번의 회귀분석을 반복함으로써 오차를 포함하지 않는 해를 구할 수 있다는 것을 수학적으로 증명하고, 상관도가 매우 높은 데이터를 이용하여 이 알고리듬에 의해 우리가 원하는 실제의 해를 구할 수 있음을 실증적으로 보여준다.

- **Theorem 1.**

$A = X'X$ 그리고 통계 툴을 이용한 $(X'X)^{-1}X'Y$의 결과를 $\widehat{\beta}$, 컴퓨터 프로그램 $(X'X+kI)^{-1}X'Y$의 결과를 B^0이라고 하고, 해의 초기 추측값을 B^0로 한다. 그 다음으로 아래에 주어진 바와 같은 방식으로 B^{m-1}을 이용하여 다음 단계에서 B^m을 구하는, 다음과 같은 반복법을 이용한다.

$$\begin{aligned} B^1 &= (X'X+kI)^{-1}(X'Y+kB^0) \\ B^m &= (X'X+kI)^{-1}(X'Y+kB^{m-1}) \quad (8) \\ m &= 1,2,3,\cdots \end{aligned}$$

이때, 이 반복법에 의해 구해진 B^m은 우리가 원하는 결과 값 $\widehat{\beta} = (X'X)^{-1}X'Y$에 수렴한다.

(증명) 정의된 반복법에 의해 다음과 같은 식을 도출한다.

$$(X'X+kI)B^m = X'Y+kB^{m-1} \quad (9)$$

$X'X\widehat{\beta} = X'Y$ 로 부터 다음의 식 (10)을 도출한다.

$$(X'X+kI)\widehat{\beta} = X'Y+k\widehat{\beta} \quad (10)$$

식 (9)에서 식 (10)을 빼면

$$(X'X+kI)(B^m-\widehat{\beta}) = k(B^{m-1}-\widehat{\beta}) \quad (11)$$

를 얻는다.

각각의 m번째 반복단계에서의 B^m의 $\widehat{\beta}$에 대한 오차를 $E^m = \|B^m-\widehat{\beta}\|$라고 할 때,

식 (6)과 식 (11)로부터 다음의 관계식을 얻는다.

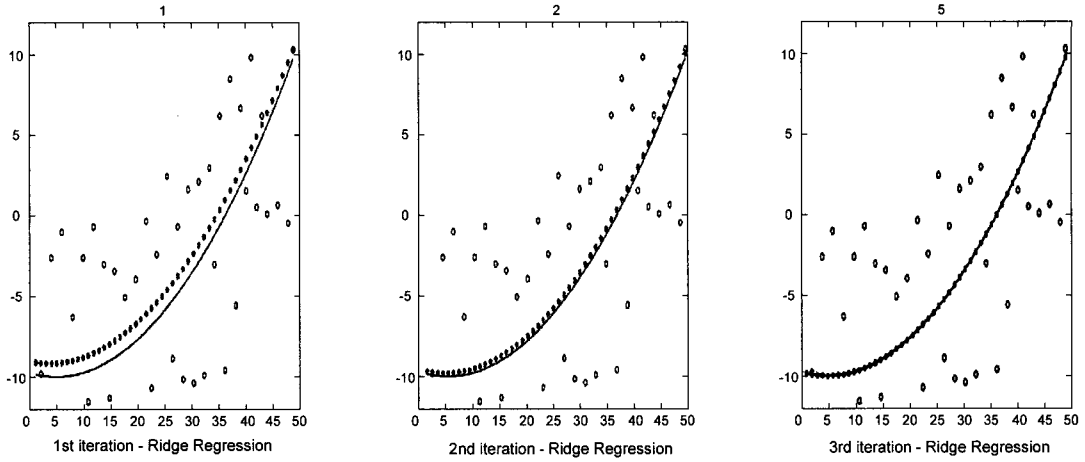
$$E^m < \max_{j=1,\cdots,n}(|k|/|\lambda_j+k|)E^{m-1} \quad (12)$$

여기서, $\widehat{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}, \cdots, \widehat{\beta_n})'$라고 할 때, 오차 $E^m = \|B^m-\widehat{\beta}\|$는 $(\sum_{j=1}^{n}|B_j^m-\widehat{\beta}_j|^2)^{1/2}$를 의미한다.

이때,

$$\max_{j=1,\cdots,n}(|k|/|\lambda_j+k|) < 1-\alpha \quad (13)$$

$\alpha > 0$ 그리고 $|1-\alpha| < 1$ 이므로 식 (12)와 식 (13)으로 부터, 각 단계에서의 E^m은 $E^m < (1-\alpha)^m E^0$의 관계식을 만족하며, 따라서, 구한 값 B^m는 빠른 속도로 $\widehat{\beta}$ 값에 가까워진다[4].



● **비교분석**

다음은 위에서 제시한 알고리듬을 다음의 regression 모델에 적용하여 보았다.

$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3^2 + \epsilon.$$

이때, X_1과 X_2는 상관도(correlation)가 매우 크도록 하였다. 또한 실질적인 데이터의 특성을 갖도록 하기 위해서, 데이터에는 난수 생성원(random number generator)을 이용하여 정규분포를 갖는 오차를 더하였다.

프로그램은 Matlab을 이용하여 실행되었다.

위의 그림에서 실선은 데이터에서 도출될 수 있는 실제의 해를 나타내며 각 점선은 프로그램을 이용하여 구한 해를 나타낸다. 그림은 각각 첫 번째, 두 번째 그리고 세 번째 반복법(iteration)에 의해 구해진 해 이다. 그림에서 첫 번째 반복법에 의해 구한 해는 Ridge regression에 의한 해이며 실제 해와는 약간의 오차를 보이고 있다. 그림에서 보듯이 세 번째 반복법, 즉 Ridge regression을 이용하여 얻은 결과로부터 두 번의 반복법을 적용한 후의 해는 실제 해와 일치함을 보여준다.

4. 결 론

회귀분석을 할 때, 만일 두 개 이상의 독립변수를 나타내는 데이터가 강한 상관관계를 가지게 되면, 회귀분석이 올바로 이루어지지 않게 된다. 이러한 경우의 회귀분석을 위한 방법이 통계학에서는 오랜 동안 연구되었으며 그 중 가장 잘 알려진 방법 중 의 하나가 Ridge regression 방법이다. 그러나 이 Ridge regression에 의한 방법은 여전히 실제의 해를 구할 수 없다는 결정적인 단점을 가지고 있다. 이에 본 연구는 수학적 알고리듬을 통해 이 단점을 보완하여 몇 단계의 Ridge regression의 반복을 통해 오차를 포함하지 않는 회귀분석결과를 도출 할 수 있도록 하였다. 이 연구결과는 위의 두 장을 통해 수학적으로 증명되고, 실제로 프로그램을 작성하여 얻은 실험결과를 제시하였다. 우리는 이러한 수학적인 알고리듬의 적용을 통해 앞으로도 데이터 분석에 있어서의 어려운 점을 보완하고 극복할 수 있을 것으로 보인다.

참 고 문 헌

[1] 장남식, 홍성완, 장재호, 「데이터 마이닝」, 대청미디어, 1999.

[2] Conte, S.D. Conte and C.D. Boor, *Elementary Numerical Analysis, An Alogrithm Approach*, (3rd ed.), McGraw-Hill, Inc., 1980.

[3] Hoerl, A.E. and R.W. Kennard, "Ridge Regression : Based Estimation for Nonorth-

ogonal Problems," *Technometrics*, Vol.12, No.1(1970), pp.55-67.

[4] Kim, S and M. Lee, "Artificial damping techniques for scalar waves in the frequency domain," *Computers Math. Applic*, Vol. 31, No.8(1996), pp.1-12.

[5] Lin, W.T. and B.B.M. Shao, "The relationship between user participation and system sucess : a simulation contingency approach," *Information & Management*, Vol.37(2000), pp.283-295.

[6] Noble, B and J.W. Daniel, *Applied Linear Algebra*, (3rd ed.) Prentice-Hall International, Inc. 1988.

[7] Seo, H.S., "A dynamic plot for the specification of curvature in linear regression," *Comp.Stat.& Data Analysis*, Vol.30(1999), pp.221-228.