응용통계연구 제 **20**권 *1*호, *2007*년, *pp. 195–204*

다중회귀모형의 그래픽적 방법

이우리[1)] 홍종선[2)] 이의기[3)]

요 약

기하학적인 방법을 사용하여 다중회귀모형 자료를 그래프로 구현하는 회귀제곱합 그림을 제안한다. 두 설명변수의 회귀제곱합은 한 변수의 단순회귀제곱합과 한 변수의 회귀모형에 다른 변수가 추가되었을 때 회귀제곱합의 증가분의 합으로 표현되는 관계식을 이용하여 회귀제곱합 그림을 반원의 형태로 구현한다. 회귀제곱합 그림은 설명변수에 대응하는 벡터로 표현되고, 반응변수에 영향력 정도를 시각적으로 구현하는 그래픽적인 방법이다. 수평축에 가까운 벡터에 대응하는 설명변수가 반응변수에 더 많은 영향을 주는 설명변수라고 판단할 수 있다. 또한 두개의 설명변수에 대응하는 벡터 사이의 각도 크기로 서프레션의 발생여부를 진단 가능하다.

주요용어: 기하학, 상관관계, 서프레션, 결정계수, 회귀제곱합.

1. 서론

다중회귀모형을 설명하는 방법으로 기하학적인 표현은 매우 유용하다. 이런 기하학적인 방법은 Box 등(1978), Margolis(1979), Herr(1980), Draper와 Smith(1981), 그리고 Bryant(1984) 등 많은 학자에 의해 연구되었다. 특히 다중회귀모형의 기하학적인 표현 방법은 서프레션(suppression)을 설명하는데 중요한 역할을 하며, Hamilton(1987, 1988)과 이에 대하여 토론한 Mitra(1988)와 Freund(1988)에 의해 발표되었다. 그리고 Schey(1993)는 서프레션을 기하학적으로 보다 많은 연구를 하였다. 앞에서 언급한 문헌들의 기하학적인 방법은 회귀제곱합(sum of squares for regression)으로 설명된다. 즉 $SSR(X_i)$, 설명변수가 하나인 X_i의 단순회귀모형에 대한 회귀제곱합, $SSR(X_iX_j)$, X_i와 X_j의 다중회귀모형에 대한 회귀제곱합, 그리고 $SSR(X_j|X_i)$, X_i가 포함된 회귀모형에 X_j가 추가되었을 때 회귀제곱합의 증가분을 바탕으로 정의된 서프레션의 관계를 상관계수로 변환하여 설명하고 이런 관계를 그래프로 구현한 Sharpe와 Roberts(1997)와 이를 확장한 Friedman과 Wall(2005) 그리고 이에 대하여 토론한 Christensen(2006)의 연구가 있다.

1) (443-760) 경기도 수원시 팔달구 이의동 산94-6, 경기대학교 응용통계학과, 교수
E-mail: wrlee@kyonggi.ac.kr
2) (교신저자)(110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
3) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 응용통계연구소, 연구원
E-mail: uikis@skku.edu

다변량 자료에서 변수들의 공분산 행렬을 이용하여 변수들의 관계를 그래픽적 방법으로 구현한 'h-plot'을 Gabriel(1971)과 Corsten과 Gabriel(1976)이 제안하였다. 그리고 Trosset(2005)은 h-plot을 발전시켜 상관계수 행렬을 바탕으로 'Correlation Diagram'을 제안하여 변수들의 상관관계를 그래픽적인 방법으로 설명하였다.

Trosset(2005)의 연구에서는 두 변수의 상관계수를 코사인 함수 $\cos(\cdot)$로 나타내는 것을 기반으로 하였고, 다중회귀모형을 기하학적으로 구현한 문헌 중 Schey(1993) 등의 연구에서도 회귀제곱합들의 관계를 코사인 함수로 표현하였다. 본 연구에서는 상관계수를 기반으로 하는 Trosset의 Correation Diagram에서 상관계수 대신에 회귀제곱합 즉, $SSR(X_i)$과 $SSR(X_iX_j)$을 바탕으로 다중회귀모형의 반응변수에 영향을 주는 설명변수들의 관계를 설명하는 그래픽적인 방법을 제안한다. 이 방법은 다중회귀모형을 구성하는 설명변수들 간의 관계를 그래픽적으로 표현하며, 특히 설명변수가 두 개인 모형에서 서프레션의 발생 정도를 앞에서 언급한 문헌에서 설명하는 방법 이외의 기하학적인 방법으로 설명할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서는 다중회귀에서 회귀제곱합을 바탕으로 제안한 그래픽적 방법인 '회귀제곱합 그림(SSR plot)'을 설명한다. 회귀제곱합 그림을 이용하여 다중회귀모형을 구현하면, 반응변수에 미치는 영향력의 정도에 따라 여러 설명변수들을 군집화할 수 있다. 그리고 앞에서 언급한 많은 문헌에서 다중회귀모형의 서프레션을 기하학적으로 설명하였는데, 본 연구에서 제안한 회귀제곱합 그림을 이용하여 변수에 대응하는 벡터 사이의 각도 크기로 서프레션의 발생여부를 설명할 수 있음을 3절에서 토론하고, 4절에서 결론을 유도한다.

2. 회귀제곱합 그림

Gabriel(1971)과 Corsten과 Gabriel(1976)의 h-plot을 발전시킨 Trosset(2005)의 Correlation Diagram은 $p \times p$ 상관계수 행렬을 시각화시킨 그림으로, 각 변수는 반지름이 1인 둥근 원에 p개의 벡터로 표현되며 변수들 사이의 각도는 변수들 간의 상관계수에 관한 정보를 포함한다. Correlation Diagram의 모든 벡터는 중심에서 원주까지를 나타내기 때문에 길이는 일정하며, 변수들 간의 관련성은 각 변수를 의미하는 벡터의 방향을 결정하는 각도 $\theta_1, \theta_2, \ldots, \theta_p$로 표현된다. 두 벡터 사이의 각도 $\theta_i - \theta_j$는 상관계수에 의존하기 때문에 모든 i와 j에 대하여 $r_{ij} \approx \cos(\theta_i - \theta_j)$ 를 만족하는 $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$를 구한다. 이를 구하기 위해 다음과 같은 최적화 문제의 형태로 해결하였다.

$$\min 2 \sum_{i<j} \left[r_{ij} - \cos(\theta_i - \theta_j) \right]^2 . \tag{2.1}$$

Trosset(2005)은 식 (2.1)의 최적화 문제를 풀기 위하여 Gay(1983, 1984)에 의해 S-Plus 함수로 개발된 준-뉴톤 알고리즘(quasi-Newton algorithm)인 *nlminb*을 사용하였다.

설명변수가 p개인 다중회귀모형을 고려하자. 설명변수 X_i와 X_j에 대한 회귀제곱합에 대한 다음과 같은 등식으로부터

$$SSR(X_iX_j) = SSR(X_i) + SSR(X_j|X_i),$$

다중회귀모형을 기하학적으로 설명한 문헌에서와 같이 다음의 관계식을 유도할 수 있다.

$$\cos^2(\theta_i - \theta_j) = \frac{SSR(X_i)}{SSR(X_iX_j)}. \tag{2.2}$$

식 (2.2)의 값이 증가할수록 $SSR(X_j|X_i)$ 값은 감소하고, 각도 $|\theta_i - \theta_j|$의 값도 감소한다. 따라서 $|\theta_i - \theta_j|$의 값이 증가하여 두 변수에 대응하는 벡터 사이의 각도가 커지면, X_i가 포함된 회귀모형에 X_j가 추가되었을 때 회귀제곱합이 증가하는 정도가 커진다는 것을 의미한다.

본 연구에서는 Correlation Diagram의 상관계수 행렬과 유사한 역할을 하는 행렬, 즉 (i,j)칸에 $SSR(X_i)/SSR(X_iX_j)$을 대입한 '회귀제곱합 행렬(SSR matrix)'을 이용한다. 회귀제곱합 행렬을 바탕으로 다음과 같은 최적화 문제를 설정하고

$$\min \sum_{i=1}^{p}\sum_{j=1}^{p}\left[\cos(\theta_i - \theta_j) - \sqrt{\frac{SSR(X_i)}{SSR(X_iX_j)}}\right]^2, \tag{2.3}$$

이 문제를 Correlation Diagram을 구현하는 방법과 동일하게 S-Plus 함수인 *nlminb* 을 사용하여 $\Theta = (\theta_1, \theta_2, \theta_3, \ldots, \theta_p)$를 구하고 회귀제곱합 그림을 작성한다.

Trosset(2005)의 Correlation Diagram에서는 대각선에 대칭인 상관계수 행렬을 사용하기 때문에 (i,j)칸과 (j,i)칸 값에 대응하는 $\cos(\theta_i - \theta_j)$와 $\cos(\theta_j - \theta_i)$는 당연히 동일하다. 다중회귀모형에서는 $\cos(\theta_j - \theta_i) = SSR(X_j)/SSR(X_iX_j)$이 되므로 $\cos(\theta_i - \theta_j)$와 $\cos(\theta_j - \theta_i)$는 서로 다른 값을 갖는다. 그러나 회귀제곱합 행렬의 (i,j)칸과 (j,i)칸에 대하여는 다음과 같은 식이 성립되어 일정한 상수 값을 갖는다.

$$\cos^2(\theta_i - \theta_j) + \cos^2(\theta_j - \theta_i) = \frac{SSR(X_i) + SSR(X_j)}{SSR(X_iX_j)}.$$

따라서 $|\theta_i - \theta_j|$의 값을 구하는 (2.3)과 같은 최적화 식의 해는 항상 존재하므로 $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$를 구할 수 있다. 그리고 설명변수 X_i와 X_j가 독립이라면, $SSR(X_j|X_i) = SSR(X_j)$이 되고 $SSR(X_iX_j) = SSR(X_i) + SSR(X_j)$이 성립한다. 따라서 $\cos^2(\theta_i - \theta_j) + \cos^2(\theta_j - \theta_i) = 1$ 이 되며, 두 개의 설명변수 X_i와 X_j만으로 구성된 회귀모형에 대한 회귀제곱합 그림에서 $|\theta_i - \theta_j| = \pi/4$이다. 즉 독립적인 두 설명변수만을 포함하는 회귀모형에서 두 변수에 대응하는 벡터 사이의 각도는 45°로 나타나게 된다. 설명변수 사이의 독립적인 관계 이외에 대하여는 3절에서 토론하기로 한다.

Correlation Diagram은 첫번째 변수에 대응하는 θ_1의 값을 0으로 설정하여 기준을 삼았지만, 회귀제곱합 그림에서는 반응변수 Y에 가장 영향력있는 설명변수에 대응하는 벡터의 각도를 0으로 간주한다. 예를 들어, X_i가 반응변수와 상관계수의 값이 제일 큰 설명변수이거나 단순회귀모형 중 가장 큰 결정계수값을 나타낸다면, 이에 대응하는 θ_i를 0으로 설정하여 수평축과 일치시킨다. 또한 상관계수는 +1에서부터 −1의 값을 나타내기 때문에 Correlation Diagram은 둥근 원으로 구현된다. 그러나 회귀제곱합 행렬의 값들은 모두 양의 값을 갖기 때문에(표 2.1과 표 2.2 참조), 각 변수에 대응하는 벡터의 각도를 의미하는 코사인 값은 $+\pi/2$부터 $-\pi/2$까지의 값을 갖는다. 따라서 회귀제곱합 그림은 오른쪽 반원의 모

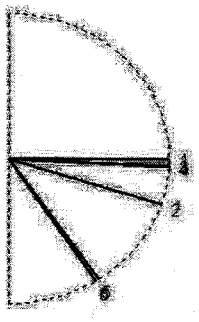

그림 2.1: 회귀제곱합 그림1

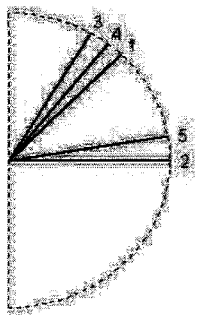

그림 2.2: 회귀제곱합 그림2

양으로 구현된다. 그리고 반응변수 Y에 가장 영향력있는 설명변수에 대응하는 벡터의 각도를 0으로 간주하기 때문에 실제로 표현되는 회귀제곱합 그림은 오른쪽 반원 중에서 위의 사분원 또는 아래의 사분원 중 하나의 형태로 구현된다(그림 2.1과 2.2 참조).

회귀분석 교재에서 예제를 발췌하고 회귀제곱합 그림을 구현하여 그림 2.1과 그림 2.2에 나타내었다. 우선 그림 2.1은 Chatterjee 등(2000, p. 53)에서 다룬 설명변수가 6개인 감독자 성과 자료(supervisor performance data)를 바탕으로 구현하였다. 반응변수와 상관계수의 값이 제일 큰 설명변수가 첫번째 변수이기 때문에 $\theta_1 = 0$으로 설정하였다. 그림 2.2는 설명변수가 5개인 Rawlings 등(1998, p. 163)의 토양의 특성 자료(soil characteristic data)를 바탕으로 구현한 회귀제곱합 그림이다. 여기서는 두번째 설명변수가 반응변수와 가장 큰 상관관계를 나타내고 있으므로 $\theta_2 = 0$으로 설정하였다. 회귀제곱합 그림을 작성하기 위한 회귀제곱합 행렬은 표 2.1과 표 2.2에 제시하였다.

그림 2.1을 살펴보면 6개의 설명변수를 두개 또는 세개의 군집으로 분류할 수 있는데 여기에서는 세개의 군집으로 분류하여 보자. 설명변수 X_1, X_3, X_4에 대응하는 벡터들로 이루어진 군집과 하나의 변수 X_2에 대한 벡터의 군집 그리고 X_5, X_6에 대응하는 벡터들로 이루어진 군집으로 분류할 수 있다. 수평축에 가깝게 군집된 벡터들에 대응하는 설명변수 X_1, X_3, X_4 들은 반응변수를 잘 설명하는 변수이고, 나머지 설명변수들은 반응변수에 커다란 영향을 주지 않는다고 결론내릴 수 있다. 이 자료에 대하여 최적화 해를 구하면서 얻은

표 2.1: 감독자 성과 자료의 회귀제곱합 행렬

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1	0.2658	0.5494	0.5092	0.0359	0.0353
X_2	0.9974	1	0.9545	0.9129	0.1284	0.1324
X_3	0.9623	0.4456	1	0.7728	0.0618	0.0556
X_4	0.9962	0.4759	0.8631	1	0.0693	0.0603
X_5	0.9999	0.9529	0.9819	0.9856	1	0.6361
X_6	0.9986	0.9995	0.8996	0.8736	0.6472	1

표 2.2: 토양의 특성 자료의 회귀제곱합 행렬

	X_1	X_2	X_3	X_4	X_5
X_1	1	0.9933	0.7840	0.9539	0.7055
X_2	0.0176	1	0.0647	0.1124	0.6410
X_3	0.1993	0.9256	1	0.9957	0.9390
X_4	0.9539	0.9103	0.5634	1	0.9066
X_5	0.7055	0.9853	0.1009	0.1721	1

식 (2.3)의 값(objective value)은 2.7440 이다. 그리고 그림 2.2에서는 5개의 설명변수들 중 X_2, X_5와 X_1, X_3, X_4에 대응하는 벡터들로 이루어진 두 개의 군집으로 형성되는 것을 파악할 수 있다. 여기에서도 수평축에 가까운 군집에 포함된 X_2, X_5 변수는 반응변수를 잘 설명하는 설명변수이고, 나머지 세개의 설명변수들은 반응변수에 영향을 미치지 않는다고 결론내릴 수 있다. 이 자료에 대한 최적해의 값은 1.6131 이다.

3. 서프레션

다중회귀모형에서 하나의 설명변수가 회귀모형에 추가되었을 경우, 다른 설명변수의 중요성을 증가시켜주는 역할을 함으로써 회귀모형의 설명력을 높여주는 변수를 서프레서(suppressor) 변수라고 하며(Hamilton, 1987), 이런 현상을 서프레션(suppression)이라고 정의한다(Horst, 1941; Conger, 1974; Cohen과 Cohen, 1975; Velicer, 1978; Friedman과 Wall, 2005). 서프레서 변수는 반응변수와는 상관관계가 적지만 다른 설명변수와는 유의한 연관성이 존재한다. 회귀분석에서는 이 현상에 대해 추정된 회귀식에 대한 설명력 정도를 보여주는 회귀제곱합들의 관계를 사용하여 설명한다. 다중회귀모형에서 서프레서 변수가 존재하면, 다음의 조건을 만족한다고 정의하였다(Velicer, 1978; Hamilton, 1987; Schey, 1993).

$$SSR(X_2|X_1) > SSR(X_2). \tag{3.1}$$

서프레션을 기하학적으로 연구한 문헌은 서론에서 언급하였듯이 많이 찾아볼 수 있다. 여기에서는 회귀제곱합의 관계를 기하학적인 방법을 사용하여 그래프로 표현한 회귀제곱합 그림으로 서프레션을 설명하고자 한다. 우선, 식 (2.2)을 다음과 같이 표현하여

$$\cos^2(\theta_1 - \theta_2) = 1 - \frac{SSR(X_2|X_1)}{SSR(X_1X_2)},$$

이 식을 $\cos^2(\theta_2 - \theta_1)$와 합하면 다음과 같다.

$$\cos^2(\theta_1 - \theta_2) + \cos^2(\theta_2 - \theta_1) = 1 - \frac{SSR(X_2|X_1) - SSR(X_2)}{SSR(X_1X_2)}. \tag{3.2}$$

만약 식 (3.1)이 성립되어 서프레션이 발생하면, 식 (3.2)의 값은 1보다 작게 된다. 이 경우에 (2.3)의 최적화의 해를 구하면, $|\theta_1 - \theta_2|$의 값은 $\pi/4$보다 큰 값을 갖는다. 그러므로 두

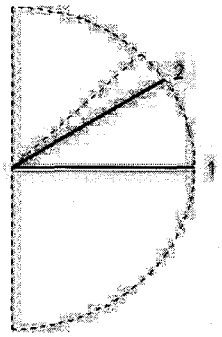

그림 3.1: 일반적 경우

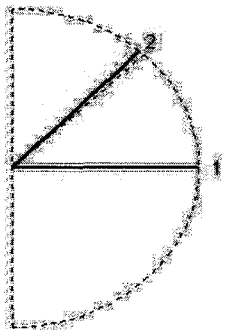

그림 3.2: 독립인 경우

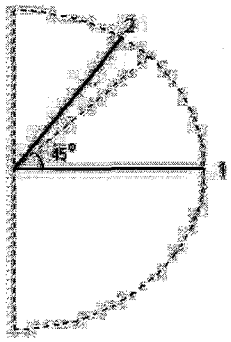

그림 3.3: 서프레션 발생

개의 설명변수의 회귀분석 자료를 회귀제곱합 그림으로 구현하여 두 변수에 대응하는 벡터 사이의 각도가 45° 이상 벌어진다면, 회귀분석에서 서프레션이 발생한다는 것을 의미한다고 결론내릴 수 있다. 또한 2절에서 설명한 바와 같이 회귀제곱합 그림에서 두 변수에 대응하는 벡터 사이의 각도가 45°에 가까운 값을 갖는다면 두 설명변수가 독립적인 관계를 갖고 있다고 판단할 수 있다.

$SSR(X_2)$와 $SSR(X_2|X_1)$과의 관계를 기하학적인 방법으로 연구한 Schey(1993)의 논문에 서프레션의 발생 여부에 따른 다양한 자료가 제시되었는데, 그 중에서 표 1의 a, b, 그리고 c의 자료를 회귀제곱합 그림으로 구현하여 그림 3.1부터 3.3에 나열하였다. 우선 그림 3.1은 Schey(1993, p. 29)의 표 1a 자료를 구현한 것으로 서프레션이 발생하지 않은 경우이다. 그림 3.1의 반원에서의 점선은 45°를 나타내는 선이고, 이때 두 벡터 사이의 각도는 45°보다 작은 33.2°이며 실선으로 표현하였다. 참고로 최적해의 값은 0.0335 이다. 식 (3.1)로부터 다음과 같이 결정계수와 상관계수로 이루어진 관계식을 유도한다(Hamilton, 1987).

$$R^2 \ > \ r_{y1}^2 + r_{y2}^2. \tag{3.3}$$

여기서 r_{y1}는 반응변수 Y와 첫번째 설명변수 사이의 상관계수이다. 결정계수와 상관계수를 구하여 보면 다음과 같이 식 (3.3)의 부등식이 성립되지 않기 때문에 서프레션이 발생하지 않는다고 파악할 수 있다.

$$R^2 = 0.653 \ < \ r_{y1}^2 + r_{y2}^2 = .780^2 + .571^2 = .934.$$

그림 3.2는 Schey의 표 1b 자료를 구현한 것으로 두 변수가 상관관계가 존재하지 않는 경우이다. 두 벡터 사이의 각도는 46.9°이고 45°에 가깝다. 따라서 2절과 3절에서 언급하였듯이 회귀제곱합 그림을 통하여 두 설명변수는 독립적이라고 식별할 수 있다. 최적해의 값은 0.0668 이며, 식 (3.3)에 결정계수와 상관계수를 대입하여 관계식을 살펴보면, 다음과 같이 등식에 수렴한다.

$$R^2 = 0.792 \ \approx \ r_{y1}^2 + r_{y2}^2 = .770^2 + .444^2 = .790.$$

따라서 두 설명변수가 독립적인 관계라고 판단할 수 있다.

Schey의 표 1c 자료를 구현한 그림 3.3은 서프레션이 발생하는 경우이다. 두 벡터 사이의 각도는 55.75°이고 이 값은 45° 보다 큰 값을 나타난다. 최적해의 값은 0.1843 이며, 결정계수와 상관계수 관계를 구하여 보면 다음과 같이 결정계수가 큰 값을 가지므로 서프레션이 발생한다고 결론내릴 수 있다.

$$R^2 = 0.9637 \quad > \quad r_{y1}^2 + r_{y2}^2 = .850^2 + .254^2 = .787.$$

그리고 서프레션에 대한 회귀제곱합 관계식 (3.1)과 상관계수의 관계식 (3.3)으로 부터 이 자료의 두번째 변수인 X_2가 서프레서 변수임을 판단할 수 있고, 회귀제곱합 그림을 통하여 서프레서 변수임을 식별할 수도 있는데, 그림 3.3에서 수평축에서 멀리 떨어진 두번째 변수는 수평축으로 나타나는 첫번째 변수보다 반응변수를 설명하는 영향력이 약하기 때문에 서프레서 변수라고 판단한다.

또한 Hamilton(1987, p. 131)의 표 1의 자료에서 $R^2 = 0.9998$이며 $r_{y1}^2 + r_{y2}^2 = .002^2 + .434^2 = .188$보다 큰 값을 갖기 때문에 서프레션이 발생하는 자료이다. 이 자료를 회귀제곱합 그림으로 구현한 결과를 살펴보면, 두 벡터 사이의 각도는 77.37°로 큰 각도를 나타내고 있어 서프레션이 발생한다는 것과 수평축이 두번째 변수에 대응하는 벡터이기 때문에 서프레서 변수는 첫번째 설명변수라는 것도 함께 회귀제곱합 그림을 통하여 탐색할 수 있다. 그러므로 본 연구에서 제안한 회귀제곱합 그림은 다중회귀모형에서 서프레션이 발생하는지의 여부와 서프레션이 발생할 때 서프레서 변수가 무엇인지도 식별 가능하다.

4. 결론

본 연구에서는 다중회귀모형 자료분석을 기하학적인 방법을 사용하여 그래프로 표현하는 대안적인 방법으로 회귀제곱합 그림을 제안하였다. 이 방법은 Trosset(2005)의 Correlation Diagram 구현의 이론을 응용하여 상관계수 대신에 2절에서 논의한 회귀제곱합의 특성을 이용하고, 회귀제곱합 행렬을 바탕으로 (2.3)식의 최적화 문제의 해를 구하여 오른쪽 반원에 벡터로 표현하였다. 오른쪽 반원에서 p개의 설명변수에 대응하는 벡터로 반응변수에 영향력의 정도를 그래픽적인 방법으로 표현하였는데, 수평축에 가까운 벡터에 해당되는 설명변수가 반응변수에 영향력을 많이 행사하는 설명변수임을 식별할 수 있다. 특히 설명변수가 두개인 경우의 회귀모형에서 설명변수에 대응하는 벡터 사이의 각도로 서프레션의 발생여부가 식별 가능하다. 즉 두 벡터 사이의 각도가 45° 이상이라면, 회귀모형에서 서프레션이 발생한다고 판단 가능하다.

참고문헌

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*, John Wiley & Sons, New York.

Bryant, P. (1984). Geometry, statistics, probability: variations on a common theme, *The American Statistician*, **38**, 38–48.

Chatterjee, S., Hadi, A. S. and Price, B. (2000). *Regression Analysis by Example*, 3rd ed., John Wiley & Sons, New York.

Christensen, R. (2006). Comment and reply to Friedman and Wall (2005), *The American Statistician*, **60**, 101–102.

Cohen, J. and Cohen, P. (1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, New Jersey.

Conger, A. J. (1974). A revised definition for suppressor variables: a guide to their identification and interpretation, *Educational and Psychological Measurement*, **34**, 35–46.

Corsten, L. C. A. and Gabriel, K. R. (1976). Graphical exploration in comparing variance matrices, *Biometrics*, **32**, 851–863.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed., John Wiley & Sons, New York.

Freund, R. J. (1988). When is $R^2 > r^2{}_{yx_1} + r^2{}_{yx_2}$ (Revisited), *The American Statistician*, **42**, 89–90.

Friedman, L. and Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression, *The American Statistician*, **59**, 127–136.

Gabriel, K. R. (1971). The biplot graphical display of matrices with applications to principal component analysis, *Biometrika*, **58**, 453–467.

Gay, D. M. (1983). Algorithm 611: subroutines for unconstrained minimization using a model/trust-region approach, *ACM Transactions on Mathematical Software*, **9**, 503–524.

Gay, D. M. (1984). A trust region approach to linearly constrained optimization, *In Numerical Analysis, Proceedings, Dundee 1983*, (F. A. Lootsma ed.), Springer, Berlin, 171–189.

Hamilton, D. (1987). Sometimes $R^2 > r^2{}_{yx_1} + r^2{}_{yx_2}$ correlated variables are not always redundant, *The American Statistician*, **41**, 129–132.

Hamilton, D. C. (1988). Reply to Freund and Mitra, *The American Statistician*, **42**, 90–91.

Herr, D. G. (1980). On the history of the use of geometry in the general linear model, *The American Statistician*, **34**, 43–47.

Horst, P. (1941). The role of prediction variables which are independent of the criterion, *The Prediction of Personal Adjustment*, (P. Horst ed.), Social Science Research Council, New York, 431–436.

Margolis, M. S. (1979). Perpendicular projections and elementary statistics, *The American Statistician*, **33**, 131–135.

Mitra, S. (1988). The relationship between the multiple and the zero-order correlation coefficients, *The American Statistician*, **42**, 89.

Rawlings, J. O., Pantula, S. G. and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*, 2nd. ed, Springer-Verlag, New York.

Schey, H. M. (1993). The relationship between the magnitudes of $SSR(x_2)$ and $SSR(x_2|x_1)$: a geometric description, *The American Statistician*, **47**, 26–30.

Sharpe, N. R. and Roberts, R. A. (1997). The relationship among sums of squares, correlation coefficients, and suppression, *The American Statistician*, **51**, 46–48.

Trosset. M. W. (2005). Visualizing correlation, *Journal of Computational & Graphical Statistics*, **14**, 1–19.

Velicer, W. F. (1978). Suppressor variables and the semipartial correlation coefficient, *Educational and Psychological Measurement*, **38**, 953–958.

[2006년 8월 접수, 2006년 12월 채택]

Graphical Method for Multiple Regression Model

W. R. Lee[1)] C. S. Hong[2)] U. K. Lee[3)]

ABSTRACT

In order to represent multiple regression data, an alternative graphical method, called as SSR Plot, is proposed by using geometrical description methods. This plot uses the relation that the sum of sqaures for regression (SSR) of two explanatory variables is known as the sum of the SSR of one variable and the increase in the SSR due to the addition of other variable to the model that already contains a variable. This half circle shaped SSR plot contains vectors corresponding explanatory variables. We might conclude that some explanatory variables corresponding to vectors which locate near the horisontal axis do affect the response variable. Also, for the regression model with two explanatory variables, a magnitude of the angle between two vectors can be identified for suppression.

Keywords: Coefficient of determination, correlation, geometry, SSR, suppression.

1) Professor, Department of Applied Statistics, Kyonggi University, Suwon 443-760, Korea
E-mail: wrlee@kyonggi.ac.kr
2) (Corresponding author) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: cshong@skku.ac.kr
3) Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 110-745, Korea
E-mail: uikis@skku.edu