영향력 관측점과 공선성에 관한 연구

김 종 우*

Influential Observations and Collinearities

Chong - woo Kim

〈목 차〉

Summary

Based on effects proposed by Mason and Gunst (1985) for the detection of outlier-induced collinearity, detailed examination of these effects shows that the relation of influential observations and collinearities are being measured and this is illustrated with examples derived from a set of data.

Ⅰ. 서 론

다중선형회귀분석(Multiple Linear Regression Analysis)에서 공선성(Collinearities)은 예측변수행렬(Predict Variable Matrix)에서 하나의 변수가 다른 하나의 변수와 또는 여러 변수들과 근사적인 종속관계(Approximate Linear Dependence) 관계를 맺고 있는 것을 말한다. 회귀계수의 최소제곱 추정값에 있어서 공선성의 효과는 잘 알려져 있으며, 이러한 효과로 부터 발생하는 문제를 해결하는 방법으로 Montgomery 와 Peck (1982)이 지적한 바와 같이 원시자료(Source Data)에 의존하게 된다.

관측점이 여러개의 독립변수로 구성될 때, 이들 중 두개 이상이 매우 큰 값을 갖는 경우에 이 관측점은 이상치(Outlier)로 나타나지 않을 수 있으며, 또한 매우 큰 값이

* 제주교육대학교 수학교육과 조교수

포함된 두 예측변수열은 공선성(Collinearity)을 발생시킬 수 있다. 이런 관측점의 경우에 회귀모형에 적합시키기 위한 방법으로 고안된 편이추정(예, Ridge Regression)이나 Single-deletion 방법등은 Mason 과 Gunst(1985)가 지적한 바와 같이 효과적인 대안이 되지 못하고 있다.

본 연구에서는 하나의 관측점이 두개 이상의 독립변수를 갖을 때, 이 관측점이 이상치로 파악되지는 않으나 그를 구성하는 독립변수들 사이에서 공선성이 유발될 수 있음을 보이고, 이 관측점과 영향력 관측점(Influential Observation)과의 관계를 조사하며 이 결과를 파악하기 위한 방법을 조사하고 예를 사용하여 분석한다.

Ⅱ. 공선성의 유발

표준화된 예측변수를 사용하여 다중선형회귀모형(Mulitiple Linear Regression Model)을 다음과 같이 정의하자.

정의 1. $Y=\beta_0 1+Z\beta+\varepsilon,\ \varepsilon_i \sim N(0, I\sigma^2)$

여기서 Y : 반응변수(response variable), n×1 벡터

Z : $Z=(Z_1, Z_2, \ldots, Z_n)$, $Z'Z$가 correlation 형식인 표준화된 nonstochastic 예측변수값으로 구성된 full-column rank 행렬

β_0 : 미지수인 회귀계수로 구성된 p×1 벡터

β : 미지수 n×1 벡터

ε : $\varepsilon_i \sim (0, I\sigma^2)$인 n×1 벡터

공선성의 존재는 반드시 세력점(Leverage Point)이 원인이지는 않지만, 세력점은 공선성을 유발할 수 있다. 이러한 점을 살펴보면, 선형회귀 모형의 X행렬이 $n\times(k+1)$의 크기로 계수가 $k+1$이고 표준화되어 있을 때, X행렬은 $X=[X_i, X_i^*]$로 나누어 보자. X_i는 X행렬의 i번째 열이고, X_i^*는 X에서 X_i를 뺀 나머지로 구성된 행렬이다. 그러면 X_i를 종속변수로 하고 나머지 k개 독립변수에 대하여 회귀방정식을 적합시킬 때 얻어지는 결정계수 R_i^2을 구하면,

$$R_i^2 = \frac{X_i' X_i^* (X_i^{*\prime} X_i^*)^{-1} X_i^{*\prime} X_i}{X_i^{*\prime} X_i - (\sum_j x_{ij})^2/n}$$

이때, 여기서 i번째 열 벡터 X_i의 값중 특정값을 무한히 크게 하면 R_i^2은 1로 접근한다. 따라서 i번째 예측변수는 공선성을 띄게 되고, 마찬가지로해서 j번째 변수도 공선성을 갖게 할 수 있다. 즉, $X_i - X_j \rightarrow 0$이 되도록 할 수 있다.

Ⅲ. 영향력 관측점의 측정

하나의 관측점이 적합된 회귀방정식(fitted equation)에서 영향력 관측점인지를 파악하기 위한 여러 방법들중에 Hat Matrix의 h_{ii}, DFFITS(i), Cook's distance, Mahalanobis' distance, Andrews-Pregibon's distance 등 여러 통계량들은 h_{ii}가 1에 접근함에 따라 관측점이 영향력 관측점인지를 파악하고 있다. Stewart(1987)은 near collinearity 상태 파악의 방법 공선성지표(Collinearity Index)를 사용하여 공선성을 유발할 수 있는 예측변수를 분석하였으며, 이 공선성지표를 사용하여 Hadi와 Velleman(1987)은 공선성과 영향력 관측점의 관계를 조사하였다.

정의 2. 독립변수 행렬 X에서 j번째 관측점의 공선성지표를 다음과 같이 정의한다.

$$k_j = \| X_j \| / \| e_j \|, \quad j=1,2,\ldots,p$$

$$= \left[x_{ij}^2 + \sum_{r=i} x_{ij}^2\right]^{\frac{1}{2}} \left[\frac{p_{ij}+p_{ij(j)}}{e_{ij}^2}\right]^{\frac{1}{2}}$$

여기서

$\| \cdot \|$: Euclidean Norm

X_j : 행렬 X의 j번째 열 벡터

$X_{(j)}$: j번째 열을 제외한 X행렬

e_j : $X_{(j)}$상에 X_j의 회귀 잔차 벡터

x_{ij} : X행렬의 ij번째 원

p : X가 생성하는 공간상에 사영된 사영행렬(Projection Matrix)

$p_{(j)}$: $X_{(j)}$가 생성하는 공간상에 사영된 사영행렬(Projection Matrix)

$p_{ij}, p_{ij(j)}$: p또는 $p_{(j)}$의 j번째 대각선상의 원

이때 $p_{ij} - p_{ij(j)} = \dfrac{e_{ij}^2}{e_j^T e_j}$ 로 나타난다.

따라서 Hadi와 Velleman은 공선성과 영향력 관찰점의 관계를 파악하기 위하여

(1) 각 독립변수간의 짝이룬 산점도 작성

(2) 각 독립변수에 대한 잔차 그림

(3) 각 독립변수에서 $p_{ij(j)}$에 대응하는 $e_{ij}^2 / e_j^T e_j$ 그림을 제안하고 있다.

공선성 지표에 의한 영향력 관측점의 위치를 Hadi와 Velleman이 제시한 그림을 통하여 살펴 보면 좌측상단과 우측하단에 위치하는데, X_j가 모델에 포함되 있으면 좌측상단에 있는 점은 높은 세력값을 갖게 되고, 반면에 우측하단에 있는 점은 X_j가 모델에 포함

여부에 관계없이 높은 세력점을 갖는 것으로 나타난다. 그리고 공선성 지표로 작성된 표에서 급격한 변동을 나타내는 예측변수는 공선성을 갖고 있음을 알 수 있다.

Ⅳ. 예

Gunst과 Mason(1980)에 있는 49개국을 대상으로 하는 6가지 사회경제변수를 갖는 GNP 자료를 사용하여 영향력 관측점과 공선성의 관계를 살펴 보자.

표 1 Collinearity Indexes

예측변수	전체자료 (n=49) (ℓ_i=.0267)		가중치자료 (n=49) (ℓ_i=.1760)		홍콩과싱가폴제외된자료 (n=47) (ℓ_i=.1802)	
	v_i	VIF	v_i	VIF	v_i	VIF
INFD	.0066	1.89	-.1931	1.83	.0991	1.94
PHYS	.0340	2.70	-.3654	2.73	.5373	2.75
DENS	.7090	19.10	.5121	2.90	.4386	2.69
AGDS	-.7034	18.85	-.5341	2.97	-.3627	2.68
LIT	.0275	3.49	-.5303	3.41	.6142	3.42
HIED	.0251	1.25	.0190	1.28	-.0619	1.24

표1은 전체자료(n=49)에서 공선성의 여부를 확인하기 위하여 최소고유치(Eigenvalue)에 대한 분산팽창계수(VIF)를 구하였다. 일반적으로 VIF가 10보다 큰 경우에 공선성 관계가 있다고 보아 DENS와 AGDS 사이에 강한 공선성이 있으리라 예측할 수 있다. 예측변수들 사이 상관계수표(표 2)에서도 이러한 점을 확인할 수 있다.

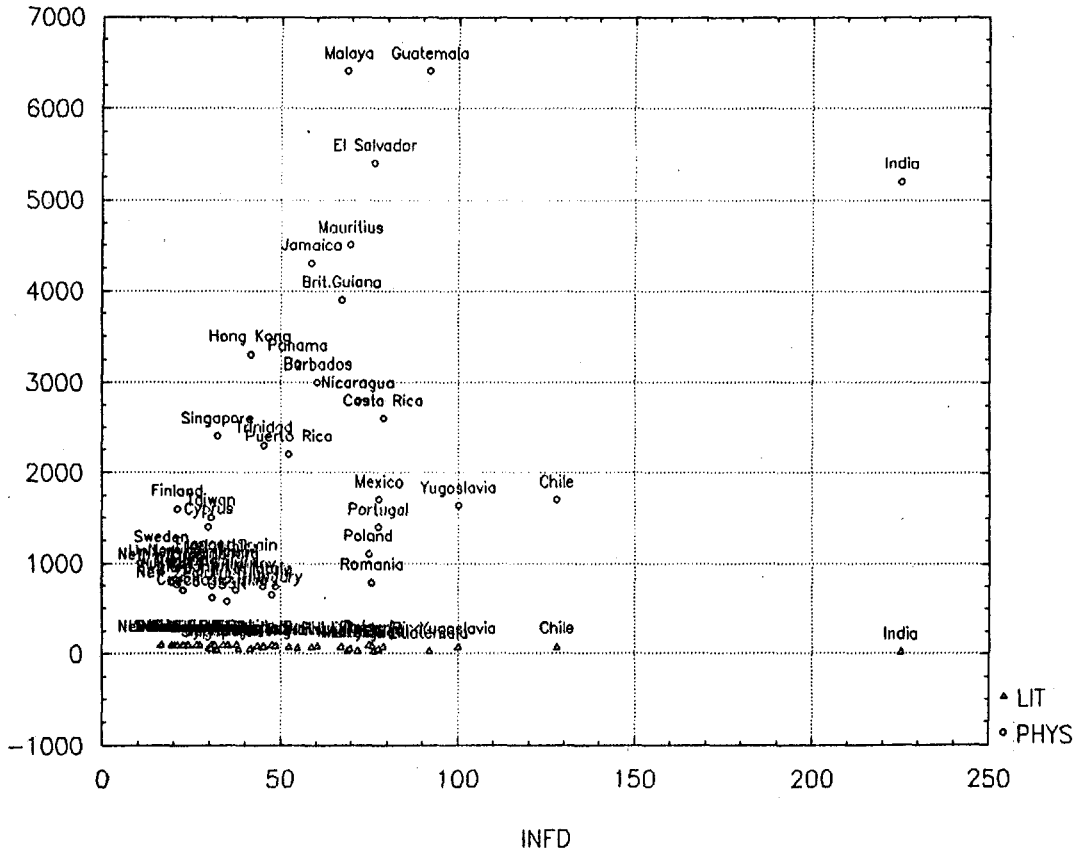
표 2 Correlation Indexes

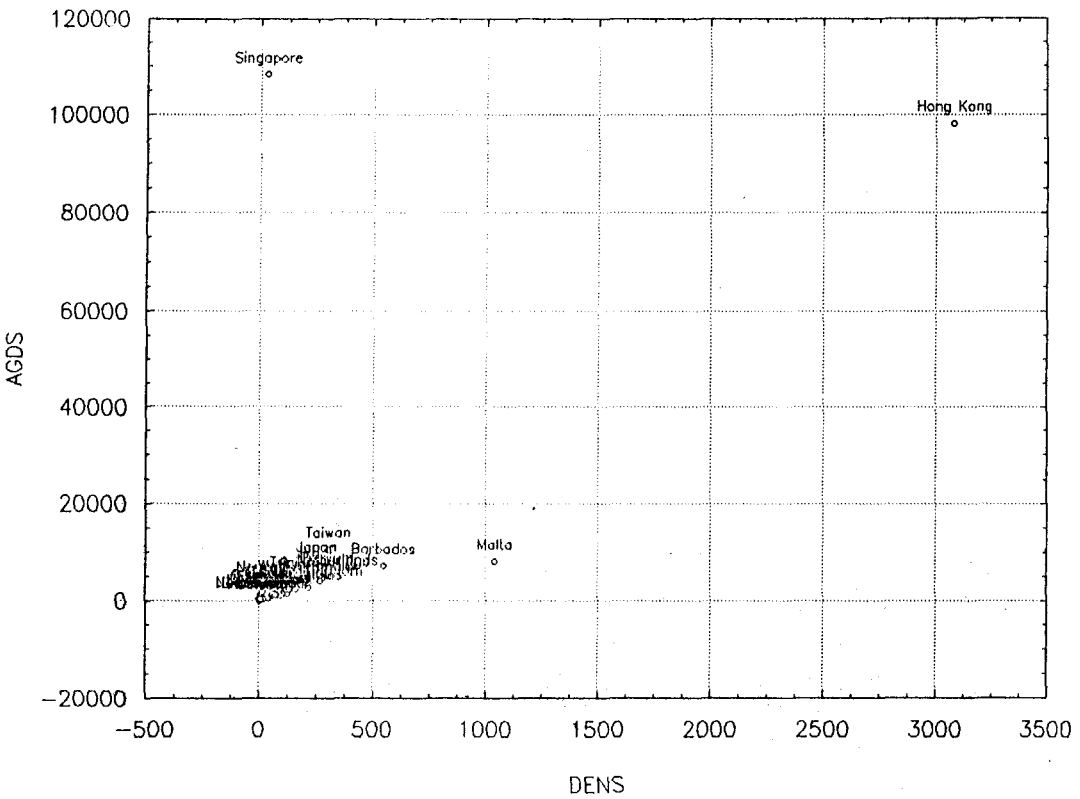
	GNP	INFD	PHYS	DENS	AGDS	LIT	HIED
GNP	1						
INFD	-.33	1					
PHYS	-.35	.57	1				
DENS	-.10	-.05	.12	1			
AGDS	-.10	-.09	.13	.62	1		
LIT	.40	-.63	-.78	-.16	-.25	1	
HIED	.32	-.31	-.37	-.14	-.07	.42	1

표 3 Leverage Value and Studentized residual

Obs.	h_{ii}	t_i	Obs.	h_{ii}	t_i
Barbados	.238	-2.026	Luxembourg	.084	2.356
Belgium	.043	1.209	Malta	.688	1.506
Canada	.042	2.011	Singapore	.632	.562
Hong Kong	.511	-.107	Taiwan	.172	-2.402
India	.558	1.337	United States	.490	.804

공선성이 이상치에 의한 것인지의 여부를 알기 위하여 이상치 파악에 일반적으로 사용되는 Belsley, Kuh, and Welsch(1980)이 제안한 스트던트화 잔차(Studentized Residual)를 표3에서 구하였다. 이 표에서 스트던트화 잔차는 어떤 관측점도 이상치로서 나타나지 않고 있음을 알 수 있으나, 관측점들 중 홍콩, 싱가폴, 말타, 인도의 세력값 (h_{ii})은 $2(p+1)/n$보다 큰 값으로 나타나 Hoaglin and Welsch(1978)의 판정 기준에





따라 영향력 관측점으로 분석할 수 있다. 이상치로서는 파악되지 않으나 세력값으로는 다수개의 영향력 관측점이 나타나는 이유는 그림에서 보여 주듯이 홍콩과 싱가폴이 서로 간에 적합한 영향을 미치는 Masking 효과를 갖기 때문이다. 또한 이런 점은 Single-deletion 진단을 사용할 경우도 동일하게 나타난다.

이상치와 영향력 관측점의 관계에서 John and Draper (1981) 는 관측점이 이상치일때 이 점이 반드시 영향력 관측점이 되지는 않는 예를 보이고 있다. 그러나 영향력을 측정하는 측도(Measure)에서 DIFFITS (i), D (i), AP (i), COVRATIO (i), FVARATIO (i) 등은 모두 추정된 회귀 계수들의 변화를 검출해내는 방법으로 만들어져서 영향을 크게 주는 관측점을 찾는 측도로 사용하나, 이들도 모두 잔차 e_i의 함수로서 나타난다. 박성현(1991)에 따르면 이들 측도에 의하여 영향을 크게 주는 관측점으로 판정될 때에 이상치가 될 가능성은 높게 나타난다고 볼 수 있다.

그림에서 보이듯이 홍콩과 싱가폴은 다른 자료의 집단에서 떨어져 있는 이상치로서 파악되어야 하고 인도와 말타는 이상치는 아니되 영향력 관측점으로 분석된다. 홍콩과 싱가폴을 제외한 적합된 회귀방정식의 계수는 전체 자료에 대한 회귀계수와 큰 차이를 보임을

다음 표4를 통하여 볼 수 있다.

표 4 Comparison of Least Squares and M Estimator

예측변수	Least Squares		M Estimator	
	n=49	n=47	n=49	n=47
INFD	-1.870	-2.076	-1.894	-2.057
PHYS	.171	.335	.119	.345
DENS	-1.094	.622	-.963	.636
AGDS	.862	-1.447	.745	-1.454
LIT	2.298	2.204	2.144	2.134
HIED	1.454	1.396	1.564	1.474

V. 결 론

관측점이 이상치로서 표출되지 않고 독립변수간에 강한 공선성을 띌 경우에 최소제곱계수에 심각한 외곡 현상을 가져 올 수 있음을 보았다. 공선성이 존재하는 자료에서 널리 사용되고 있는 Single-deletion 방법이나 편이 추정방법등은 이러한 관측점의 제어에 실패하고 있다. 따라서 VIF나 높은 h_{ii}의 확인, 상관관계의 조사 또는 공선성 지표를 사용한 공선성 징후가 있는 예측변수들을 조사하고 이들을 산점도를 통하여 확인한 후에 회귀추정을 실시하여야 한다

그렇지만 오늘날 어떠한 방법도 단독으로 공선성과 이상치, 영향력 관측점을 자동으로 보안해 주지는 못하고 있다. 몇가지 진단 방법이나 대안으로 사용되는 추정기법을 사용하여 심도 깊게 분석하여 얻어진 정보는 보다 적절한 해결 방법이라 여겨진다.

참 고 문 헌

박성현, 회귀분석, 박영사, 1991.

Andrews, D.F., and Pregibon, D. (1978), "Finding the Outliers that Matter", The Journal of the Royal Statistical Society Series-B, 40, 85-91.

Atkinson, A.C. (1981), "Two graphical displays for outlying and influential observations in regression", Biometrica, 68, 13-20.

Atkinson, A.C. (1985), Plots, Transformations and Regression, Oxford. U.K. : Oxford University Press.

Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression", Technometrics 19, 15-18.

Draper, N.R., and John, J.A. (1981), "Influential Observations and Outliers in Regression", Technometrics 23, 21-26.

Draper, N.R., and Smith, H. (1966), Applied Regression Analysis, New York : John Wiley & Sons.

Gunst, R.F., and Mason, R.L. (1980), Regression Analysis and its Application, New York : Marcel Dekker.

Hadi, A.S., and Velleman, P.F. (1987), Comment on "Collinearity and Least Squares Regression", by G.W. Stewart, Statistical Science 2, 68-100.

Hoaglin, A.E., and Welsch, R.E (1978), "The Hat Matrix in Regression and ANOVA", The American Statistician, 32, 17-22.

Marasinghe, M.G. (1985), "A Multistage Procedure for Detecting Several Outlier in Linear Regression", Technometrics 27, 395-407.

Mason, R.L., and Gunst, R.F. (1985), "Outlier-Induced Collinearitie", Technometrics 27, 401-407.

Montgomery, D.C., and Welsch, R.E. (1982), Introduction to Linear Regression Analysis, New York : John Wiley.

Stewart, G.W. (1987), "Collinearity and Least Squares Regression (with discussion)", Statistical Science 2, 68-100.