Communications of the Korean Statistical Society Vol. 15, No. 4, 2008, pp. 509–516

변이계수에 대한 영향함수

이윤희[1], 김홍기[2]

요 약

본 논문에서는 변이계수에 대한 영향함수를 유도한다. 경험적 영향함수와 표본 영향함수를 이용하여 유도된 영향함수의 타당성을 입증하고 이를 위하여 정규분포 $N(20, 1^2)$와 $N(20, 5^2)$에서 각각 확률표본을 추출하여 시뮬레이션을 수행한다. 시뮬레이션 결과로부터, 유도된 변이계수에 대한 영향함수가 한 개의 관찰치가 제거되었을 때 변이계수의 변화량을 매우 정확히 추정하는 것을 확인하였다.

주요용어: 영향함수; 변이계수.

1. 서론

주어진 자료 중에서 다른 관찰치에 비해 유난히 적거나 큰 값으로 정의되는 이상치(outlier)는 자료로부터 계산되는 통계량에 큰 영향을 미쳐 엉뚱한 결론을 유도할 수 있으므로 사전에 제거되어야 한다. 영향함수는 연속함수에서의 일차미분계수와 같은 개념으로, 우리가 $f'(x)\Delta x$로 $f(x+\Delta x)-f(x)$의 값을 근사적으로 구하고 예측하듯이, 여러 통계량에서 한 개의 관찰치가 더해지거나 빼질 때(주로 빼질 때)의 영향(변화)을 계산할 수 있도록 해준다. 통계량에 큰 영향을 미치는 값이 이상치일 가능성이 매우 크므로 손쉬운 영향함수의 계산을 통해 이상치를 확인할 수 있다.

영향함수는 Hampel (1974)에 의하여 처음으로 소개되어 지금까지 많은 연구가 진행되고 있다. Hampel (1974)은 영향함수가 모수 및 통계량에 각각 적용 가능함을 보였고, Campbell (1978)은 판별분석(Discriminant analysis)에 영향함수를 적용하여 이상치를 탐지하였으며, Radhakrishnan과 Kshirsagar (1981)은 다변량 분석에서 여러 가지 모수에 대한 이론적인 영향함수를 유도하였다. 또한 Cook (1977), Cook과 Weisberg (1980, 1982)는 회귀분석에서 회귀진단법으로, Critchley (1985)는 주성분 분석에서 영향력 있는 관찰치를 찾기 위해 이 방법을 적용하였다. Kim (1992)은 이차원 분할표의 대응분석에서 고유치들에 대한 영향함수를 유도하였으며 이를 다차원 분할표의 대응분석으로 확장하여 적용하였다. Kim과 Lee (1996), Kim (1998)은 χ^2 통계량에 대한 영향함수를 다루었고, Kim 등 (2003)은 허용한계에 대한 영향함수를 그리고 Kim과 Kim (2005)은 t 통계량에 대한 영향함수를 유도하였다.

1) (305-764) 대전광역시 유성구 궁동 220 충 남대학교 정보통계학과, 박사. E-mail: yunhlee@cnu.ac.kr
2) (305-764) 대전광역시 유성구 궁동 220 충남대학교 정보통계학과, 교수.
교신저자: honggiekim@cnu.ac.kr

본 논문에서는 표본 집단의 상대적 변이성을 측정하는 변이계수에 대하여 각 관찰치가 갖는 영향력 측정함수를 유도하고 이를 평가한다. 2절에서는 영향함수 및 변이계수의 특성을 기술하고, 영향함수의 정의를 이용하여 변이계수에 대한 영향함수를 구한다. 또한 표본으로부터 얻어지는 변이계수의 경험적 영향함수와 표본영향함수를 구한다. 3절에서는 유도된 영향함수를 검증하기 위하여 모의실험을 수행한다. 모의실험에서는 정규분포에서 뽑은 확률표본을 이용하여 경험적 영향함수와 표본영향함수가 근사적으로 동일함을 보이고 이를 통해 유도된 변이계수에 대한 영향함수의 타당성을 검증한다.

2. 변이계수와 영향함수

2.1. 영향함수

T는 분포함수에 대해 실수값(real value)을 갖는 범함수(functional)라고 하자. 즉 범함수 T는 일련의 모수이다. T의 분포함수를 $F(t)$라 하고 $\delta_x(t)$는 실수 공간의 한 점인 x에서 확률이 1인 분포함수이고 다음과 같이 정의한다.

$$\delta_x(t) = \begin{cases} 0, & t < x, \\ 1, & t \geq x, \end{cases}$$

여기서 $\delta_x(t)$를 퇴화분포함수(degenerate distribution function)라고 한다. 편의상 $F(t)$와 $\delta_x(t)$를 F와 δ_x로 쓰기로 한다.

분포함수 F와 δ_x의 혼합분포함수 F_ϵ를 다음과 같이 정의하자.

$$F_\epsilon = (1-\epsilon)\,F + \epsilon\delta_x, \quad 0 < \epsilon < 1,$$

여기서 F_ϵ을 F의 섭동(perturbation)이라 한다.

범함수 $T(F)$에 대한 x의 영향함수(influence function)는 Hampel (1974)에 의해서 다음과 같이 정의되었다.

$$\begin{aligned} \mathrm{IF}(T,\, x) &= \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{T\left[(1-\epsilon)\,F + \epsilon\delta_x\right] - T(F)}{\epsilon}. \end{aligned} \tag{2.1}$$

영향함수는 추가된 x에 의해 섭동된 범함수 $T(F_\epsilon)$의 원래의 범함수 $T(F)$로부터의 순간변화율을 나타내므로 관찰치 x에서의 $T(F)$의 영향함수값으로 $T(F)$에 대한 x의 영향을 표현할 수 있다. 한편, 식 (2.1)에서 정의한 영향함수는 l'Hospital의 정리에 의해 다음과 같이 표현할 수 있다.

$$\mathrm{IF}(T,\, x) = \left[\frac{\partial T(F_\epsilon)}{\partial \epsilon}\right]_{\epsilon=0}. \tag{2.2}$$

모집단의 평균과 분산은 범함수 $T(F)$의 일종이므로 각각을 범함수

$$\mu(F) = \mu = \int t\, dF, \qquad \sigma^2(F) = \sigma^2 = \int (t-\mu)^2\, dF$$

로 표현하며, 식 (2.1) 또는 (2.2)를 이용하여 평균과 분산에 대한 영향함수를 구하면 다음과 같이 얻을 수 있다.

$$\mathrm{IF}(\mu,\, x) = x - \mu, \tag{2.3}$$

$$\mathrm{IF}(\sigma^2,\, x) = (x-\mu)^2 - \sigma^2. \tag{2.4}$$

식 (2.3)에서 보는 바와 같이 평균에 대한 영향함수는 관찰치와 평균의 편차로 설명할 수 있다. 즉 관찰치가 평균보다 큰 값이면 평균을 높이는 양의 방향으로 영향을 주고, 평균보다 적은 값이면 평균을 낮추는 음의 방향으로 영향을 주는 것을 알 수 있다. 분산에 대한 영향함수인 식 (2.4)를 살펴보면, 관찰치가 평균 μ로부터 갖는 거리의 제곱인 $(x-\mu)^2$이 분산 σ^2과 같으면 영향함수의 값이 0이 되는데, 이는 이러한 관찰치는 분산에 대한 영향이 없다는 것을 의미한다. 한편 $(x-\mu)^2$이 σ^2보다 큰 값을 갖는 관찰치는 분산을 높이는 쪽으로 영향을 가지며, $(x-\mu)^2$이 σ^2보다 작은 값을 갖는 관찰치는 그 반대의 영향을 미친다.

Kim 등 (2003)은 분산에 대한 영향함수인 식 (2.4)를 이용하여 표준편차 σ에 대한 영향함수를 유도하였다. 표준편차는 $\sigma = \sqrt{\sigma^2}$이므로 이를 범함수 형태로 표현하면 $\sigma(F) = \sqrt{\sigma^2(F)}$로 쓸 수 있고 이것의 영향함수는 다음 식과 같이 얻을 수 있다.

$$\begin{aligned}
\mathrm{IF}(\sigma,\, x) &= \left[\frac{\partial\sqrt{\sigma^2\left(F_\epsilon\right)}}{\partial\epsilon}\right]_{\epsilon=0} \\
&= \left[\frac{1}{2}\frac{1}{\sqrt{\sigma^2\left(F_\epsilon\right)}}\frac{\partial\sigma^2\left(F_\epsilon\right)}{\partial\epsilon}\right]_{\epsilon=0} \\
&= \frac{1}{2\sigma(F)}\cdot\mathrm{IF}(\sigma^2,\, x).
\end{aligned} \tag{2.5}$$

따라서 $\mathrm{IF}(\sigma,\, x) = 1/(2\sigma)\left\{(x-\mu)^2 - \sigma^2\right\}$임을 알 수 있다.

2.2. 변이계수에 대한 영향함수

모집단의 평균이 $\mu(>0)$이고 표준편차가 σ일 때, 평균에 대한 표준편차의 비율을 변이계수(coefficient of variation)라 하고 이를 식으로 표현하면 다음과 같다.

$$V = \frac{\sigma}{\mu}. \tag{2.6}$$

식 (2.6)의 변이계수는 주로 백분율(%)로 변환하여 사용한다. 자료의 변이성을 측정할 때, 표준편차는 자료의 절대적 변이성을 측정하는데 사용되는 반면, 변이계수는 자료의 상대적 변이성을 측정하는데 사용된다. 변이계수는 자료의 단위가 다를 경우 또는 단위가 같더라도 평균의 차이가 클 경우에 유용하다. 즉 중심의 위치가 상이한 두 개 이상의 집단의 산포도를 비교할 때 효과적이다.

식 (2.6)에서 정의한 변이계수는 범함수인 평균 μ와 표준편차 σ의 함수이므로 범함수에 해당된다. 변이계수에 대한 영향함수 $\mathrm{IF}(\sigma/\mu,\, x)$는 식 (2.2)를 이용하여 다음과 같이

구할 수 있다.

$$\begin{aligned}
\mathrm{IF}\left(\frac{\sigma}{\mu}, x\right) &= \left[\frac{\partial}{\partial\epsilon}\frac{\sigma(F_\epsilon)}{\mu(F_\epsilon)}\right]_{\epsilon=0} \\
&= \left[\frac{\mu(F_\epsilon)\dfrac{\partial}{\partial\epsilon}\sigma(F_\epsilon) - \sigma(F_\epsilon)\dfrac{\partial}{\partial\epsilon}\mu(F_\epsilon)}{\mu^2(F_\epsilon)}\right]_{\epsilon=0} \\
&= \left[\frac{1}{\mu(F_\epsilon)}\frac{\partial}{\partial\epsilon}\sigma(F_\epsilon) - \frac{\sigma(F_\epsilon)}{\mu^2(F_\epsilon)}\frac{\partial}{\partial\epsilon}\mu(F_\epsilon)\right]_{\epsilon=0} \\
&= \frac{1}{\mu}\mathrm{IF}(\sigma, x) - \frac{\sigma}{\mu^2}\mathrm{IF}(\mu, x) \\
&= \frac{1}{2\mu\sigma}\left\{(x-\mu)^2 - \sigma^2\right\} - \frac{\sigma}{\mu^2}(x-\mu). \qquad (2.7)
\end{aligned}$$

식 (2.7)에서 보는 바와 같이 변이계수에 대한 영향함수는 평균 및 표준편차의 영향함수 $\mathrm{IF}(\mu, x)$와 $\mathrm{IF}(\sigma, x)$의 함수형태로 얻어짐을 알 수 있다. 평균에 대한 영향함수 값은 변이계수에 대한 영향함수의 값을 낮추는 음의 방향으로 영향을 주고, 표준편차에 대한 영향함수 값은 변이계수의 영향함수의 값을 높이는 양의 방향으로 영향을 주는 것을 알 수 있다.

한편 변이계수에 대한 영향함수는 평균 μ가 충분히 클 경우에, 즉 $\mu \gg 0$일 경우에 의미가 있다. $\mu \simeq 0$일 때는 식 (2.7)이 불완전하기 때문에 변이계수에 대한 영향함수 식 (2.7)은 $\mu \gg 0$인 경우에 사용하는 것이 타당하다.

2.3. 경험적 분포함수로부터 얻어지는 영향함수

영향함수가 모집단 분포함수 F에서의 범함수 $T(F)$에 대하여 정의된다면, 분포함수 F대신 $\hat{F}$에서의 범함수의 추정량 $T(\hat{F})$에 대한 영향함수를 경험적 영향함수(Empirical Influence Function: EIF)라고 한다. 경험적 분포함수 $\hat{F}$에 의하여 정의되는 추정량 $T(\hat{F})$에 대한 영향함수에는 세 가지가 있다. 첫 번째는 범함수의 영향함수식에서 모수 대신 모수의 추정량으로 구한 영향함수 $\mathrm{EIF}(T(\hat{F}), x_i)$이고, 두 번째는 $\hat{F}$대신 i번째 관찰치 x_i를 제거한 상태에서 구한 $\hat{F}_{(i)}$에서의 영향함수 $\mathrm{EIF}(T(\hat{F}_{(i)}), x_i)$이다. 세 번째는 분포함수 F 대신에 경험적 영향함수 $\hat{F}$을 이용하고, 섭동인 ϵ는 주어진 표본에 i번째 관찰치가 추가되거나 제거되는 경우에 대하여 각각 $1/(n+1)$과 $-1/(n-1)$으로 대신하여 얻어지는 표본영향함수(Sample Influence Function: SIF)이다.

본 논문에서는 첫 번째에 해당되는 경험적 영향함수와 표본영향함수의 정의를 이용하여 표본으로부터 구해지는 변이계수에 대한 영향함수를 구한다. 경험적 분포함수 $\hat{F}$으로부터 구해지는 변이계수의 추정량은

$$v = \frac{s}{\bar{x}}$$

이므로, 변이계수에 대한 관찰치 x_i의 경험적 영향함수는 다음 식과 같이 얻어진다.

$$\mathrm{EIF}(v, x_i) = \frac{1}{2\bar{x}s}\left\{(x_i-\bar{x})^2 - s^2\right\} - \frac{s}{\bar{x}^2}(x_i-\bar{x}) \qquad (2.8)$$

또한 변이계수에 대한 표본영향함수는 주어진 표본에서 x_i를 제거한 경우에 대하여 다음과 같이 구할 수 있다.

$$\begin{aligned} \mathrm{SIF}\,(v,\, x_i) &= -(n-1)T\left(\hat{F_{(i)}}\right) - T(\hat{F}) \\ &= -(n-1)(v_{(i)} - v), \end{aligned} \tag{2.9}$$

여기서 $v_{(i)}$는 x_i를 뺀 후의 변이계수의 값으로 $v_{(i)} = s_{(i)}/\bar{x}_{(i)}$이다.

3. 모의실험

본 논문에서 구한 변이계수에 대한 영향함수의 타당성을 검증하기 위하여 모의실험을 실시하였다. 변이계수에 대한 영향함수의 타당성은 다음 식과 같이 식 (2.8)에서의 경험적 영향함수와 식 (2.9)에서의 표본영향함수가 근사적으로 동일함을 만족할 때 입증된다.

$$\mathrm{EIF}(v,\, x_i) \simeq -(n-1)(v_{(i)} - v). \tag{3.1}$$

모의실험은 정규분포의 평균이 충분히 큰 경우에 대하여 표준편차가 적은 경우와 큰 경우로 나누어 수행하였다. 즉 정규분포 $N(20,\, 1^2)$와 $N(20,\, 5^2)$에서 각각 30개의 표본을 추출하였고 $\bar{x}$, s, $\bar{x}_{(i)}$와 $s_{(i)}$를 계산하여 v와 $v_{(i)}$를 구한 후, 식 (3.1)을 만족하는지를 알아보았다.

$N(20,\, 1^2)$에서 얻은 표본의 평균 및 표준편차는 각각 $\bar{x} = 19.9293$과 $s = 0.8129$로 변이계수는 $v = 0.0408$로 구해졌고, $N(20,\, 5^2)$인 경우에는 $\bar{x} = 19.6466$, $s = 4.0644$로부터 변이계수가 $v = 0.2069$로 구해졌다. 각각의 경우에 대한 경험적 영향함수 값과 표본영향함수 값을 표 3.1에 제시하였다. 표 3.1에서 보는 바와 같이, 두 경우 모두 모든 관찰치에서의 경험적 영향함수 값과 표본영향함수 값이 거의 동일한 값으로 나타났다.

4. 결론

본 논문에서는 평균에 대한 표준편차의 상대적 크기를 여러 집단별로 비교하는데 주로 사용하는 변이계수 v에 대한 영향함수 $\mathrm{IF}(v,\, x)$를 유도하였고 모의실험을 통하여 이 영향함수의 타당성을 검증하였다.

변이계수에 대한 영향함수는

$$\mathrm{IF}\left(\frac{\sigma}{\mu},\, x\right) = \frac{1}{2\mu\sigma}\left\{(x-\mu)^2 - \sigma^2\right\} - \frac{\sigma}{\mu^2}(x-\mu) \tag{4.1}$$

와 같이 구해졌다. 이 영향함수는 확률표본으로부터 얻어지는 경험적 영향함수와 표본영향함수가 거의 일치함을 보임으로 인해 타당성이 입증되었으므로, 변이계수에 대하여 각 관찰치가 갖는 고유한 영향력을 측정하는데 유용할 것이며 이상치를 찾는데도 기여할 것이다.

표 3.1: 모의실험을 통한 경험적 영향함수와 표본영향함수 비교

ID	$X \sim N(20, 1^2)$			ID	$X \sim N(20, 5^2)$		
	x_i	SIF(v, x_i)	EIF(v, x_i)		x_i	SIF(v, x_i)	EIF(v, x_i)
1	18.1473	0.0905	0.0813	1	10.7364	0.5273	0.4875
2	18.9299	0.0140	0.0125	2	14.6493	0.1124	0.1056
3	19.0159	0.0083	0.0072	3	15.0797	0.0803	0.0752
4	19.0229	0.0079	0.0068	4	15.1143	0.0779	0.0729
5	19.1697	−0.0005	−0.0010	5	15.8485	0.0295	0.0269
6	19.2574	−0.0048	−0.0051	6	16.2869	0.0040	0.0026
7	19.2636	−0.0051	−0.0054	7	16.3181	0.0023	0.0010
8	19.2839	−0.0060	−0.0062	8	16.4196	−0.0032	−0.0042
9	19.2920	−0.0064	−0.0066	9	16.4602	−0.0053	−0.0063
10	19.3090	−0.0071	−0.0072	10	16.5452	−0.0097	−0.0105
11	19.5804	−0.0163	−0.0159	11	17.9021	−0.0675	−0.0660
12	19.5924	−0.0165	−0.0162	12	17.9621	−0.0695	−0.0679
13	19.6227	−0.0172	−0.0169	13	18.1133	−0.0743	−0.0726
14	19.6833	−0.0185	−0.0180	14	18.4167	−0.0831	−0.0810
15	19.8524	−0.0206	−0.0201	15	19.2622	−0.1011	−0.0985
16	19.8763	−0.0207	−0.0202	16	19.3813	−0.1029	−0.1002
17	20.0180	−0.0209	−0.0203	17	20.0901	−0.1097	−0.1069
18	20.0186	−0.0209	−0.0203	18	20.0928	−0.1097	−0.1069
19	20.0484	−0.0207	−0.0202	19	20.2421	−0.1102	−0.1075
20	20.1654	−0.0196	−0.0192	20	20.8269	−0.1097	−0.1071
21	20.2334	−0.0186	−0.0182	21	21.1672	−0.1072	−0.1050
22	20.4732	−0.0124	−0.0124	22	22.3659	−0.0863	−0.0858
23	20.6431	−0.0057	−0.0061	23	23.2154	−0.0596	−0.0613
24	20.6756	−0.0042	−0.0047	24	23.3779	−0.0534	−0.0555
25	20.8300	0.0039	0.0028	25	24.1502	−0.0186	−0.0239
26	20.9582	0.0119	0.0102	26	24.7910	0.0168	0.0081
27	21.0302	0.0169	0.0148	27	25.1512	0.0394	0.0283
28	21.1556	0.0265	0.0235	28	25.7779	0.0832	0.0674
29	21.3222	0.0411	0.0366	29	26.6109	0.1509	0.1269
30	21.4085	0.0495	0.0441	30	27.0425	0.1903	0.1612

참고문헌

Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, **27**, 251–258.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

Cook, R. D. and Weisberg, S. (1980). Characterization of an empirical influence function for detecting influential cases in regression, *Technometrics*, **22**, 495–508.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall/CRC, New York.

Critchley, F. (1985). Influence in principal components analysis, *Biometrika*, **72**, 627–636.

Hample, F. R. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, **69**, 383–393.

Kim, H. (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, **40**, 201–217.

Kim, H. (1998). A study on cell influences to Chi-square statistic in contingency tables, *The Korean Communications in Statistics*, **5**, 35–42.

Kim, H. and Kim, K. H. (2005). Influence of an observation on the t-statistic, *The Korean Communications in Statistics*, **12**, 453–462.

Kim, H. and Lee, H. S. (1996). Influence functions on χ^2 statistics in contingency tables, *The Korean Communications in Statistics*, **3**, 69–76.

Kim, H., Lee, Y. H., Shin, H. S. and Lee, S. (2003). Influence function on tolerance limit, *The Korean Communications in Statistics*, **10**, 497–505.

Lee, H. and Kim, H. (2003). The changes in χ^2 statistic when a row is deleted from a contingency table, *The Korean Communications in Statistics*, **10**, 305–317.

Radhakrishnan, R. and Kshirsagar, A. M. (1981). Influence functions for certain parameters in multi-variate analysis, *Communications in Statistics - Theory and Methods*, **10**, 515–529.

[2007년 11월 접수, 2008년 4월 채택]

Influence Function on the Coefficient of Variation

Yun-Hee Lee[1], Honggie Kim[2]

Abstract

We derive the influence function on the coefficient of variation. Empirical influence function and Sample influence function are used to verify the validity of the derived influence function. To show the validity of the influence function, we carry out simulations with random samples from normal distribution $N(20, 1^2)$ and $N(20, 5^2)$, respectively. The simulation result proves that the derived influence function is very accurate in estimating changes in the coefficient of variation when an observation is deleted.

Keywords: Influence function; coefficient of variation.

1) Ph.D, Department of Information and Statistics, Chungnam National University, Daejeon 305-764, Korea. E-mail: yunhlee@cnu.ac.kr

2) Professor, Department of Information and Statistics, Chungnam National University, Daejeon 305-764, Korea. Correspondence: honggiekim@cnu.ac.kr