The Korean Journal of Applied Statistics (2017) DOI: https://doi.org/10.5351/KJAS.2017.30.4.529
30(4), 529–538

Derivation and verification of influence function on parameter δ proposed by Ghosh and Kim

Minjeong Kim[a] · Honggie Kim[b,1]

[a]Income Statistics Division, Statistics Korea;
[b]Department of Information and Statistics, Chungnam National University

(Received April 10, 2017; Revised June 26, 2017; Accepted July 27, 2017)

Abstract

The Ghosh and Kim zero-altered distribution model is used to analyze count data that have too many or too few zeros. The dispersion type parameter δ in the zero-altered distribution model consists of mean, variance and zero probability and has two forms depending on the relation between μ and σ^2. We derived the influence function on δ when $\sigma^2 \geq \mu$. To show the validity of the influence function, we used the Census data on the number of births of married women in Korea to compare the estimated changes in δ using this function with those obtained using the direct deletion method. The result proved that the obtained influence function is very accurate in estimating changes in δ when an observation is deleted.

Keywords: over-dispersion, under-dispersion, dispersion parameter, zero-altered model, influence function

1. 서론

조사 자료를 이용한 데이터 분석에서 각각의 데이터가 가지는 영향력에 대해 파악하는 것은 의미가 있다. 하나의 데이터라도 분석에 큰 영향을 미쳐 잘못된 의사결정을 내릴 수 있기 때문이다. 이렇게 자료 집합 내에서 다른 관측치들과 통계학적으로 유의하게 다른 관측치를 이상치(outlier)라 부른다. 자료에 이상치가 있을 경우 평균 등 일부 통계량은 큰 영향을 받게 되므로 자료 분석에 앞서 이상치 유무를 먼저 확인하여 이상치가 있다면 원인에 따라 알맞은 처리가 선행되어야 올바른 분석이 이루어질 수 있다.

보다 쉽고 빠르게 각각의 관측치가 통계량에 미치는 영향력을 측정하고 그 중 이상치를 찾아내는 통계학적 방법론에 대한 연구 중 하나로 1974년 Hampel이 소개한 영향함수(influence function)가 있다. 영향함수는 연속함수의 일차미분계수와 동일한 개념으로 여러 통계량에서 한 개의 관측치를 더하거나 뺄 때 통계량에 미치는 영향을 함수를 통해 추정할 수 있도록 해준다.

Hampel (1974)에 의해 영향함수가 모수 및 대부분의 통계량에 적용 가능함이 확인된 이래, Campbell (1978)이 판별분석(discriminant analysis)에 영향함수를 적용하여 이상치를 탐지하였으며, Radhakrishnan과 Kshirsagar (1981)은 다변량 분석에서 여러 모수에 대한 이론적 영향함수를 유도하였다. Cook과 Weisberg (1980, 1982)는 회귀분석에서 회귀진단방법으로, Critchley (1985)는 주성분 분석에

This research was supported by 2016 Chungnam National University Research Fund.
[1]Corresponding author: Department of Information and Statistics, Chungnam National University, 99 Daehak-ro, Yuseong-gu, Daejeon 34134, Korea. E-mail: honggiekim@cnu.ac.kr

서 영향력 있는 관찰치 탐색을 위해 영향함수를 이용하였다. 국내연구에서는 이차원 분할표의 대응분석에서 구한 고유치들에 대한 영향함수를 유도하였고 (Kim, 1992), 이를 다차원 분할표의 대응분석으로 확장하였다 (Kim, 1994). 이후에도 통계량에 대한 영향함수 (Kim과 Lee, 1996; Kim, 1998), 허용한계에 대한 영향함수 (Kim 등, 2003), t-통계량에 대한 영향함수 (Kim, 2005) 등 관련 연구가 지속적으로 진행되었다.

본 논문에서는 비음정수값을 갖는 변수에 적용되는 Ghosh와 Kim (2007)의 영 변환(zero-altered) 모형 모수 δ에 대한 영향함수를 유도하고 통계청 인구주택총조사(표본)의 '연령 및 출생자녀수별 기혼여성인구(15세 이상)' 자료를 적용하여 유도한 영향함수의 타당성을 평가하였다.

2. 영 변환 모형과 영향함수

2.1. 영 변환 모형(zero-altered model)

Ghosh와 Kim (2007)은 포아송 분포와 같이 균등산포를 가정한 모형에서 영 과잉과 영 부족을 동시에 고려한 확률 모형으로 '영 변환 모형'을 제안하였다.

공간 $N = \{0, 1, 2, \ldots\}$에서 값을 가지는 이산확률변수 U의 확률질량함수를 $f_0(u)$라 하고, 모든 u에서 $f_0(u) < 1$이라 가정하고, 이 확률분포의 $u = 0$에서의 확률 $f_0(0)$를 임의로 변경했을 때 Ghosh와 Kim (2007)의 산포형태모수(dispersion type parameter) $\delta \in (-1, 1)$가 정의되면 다음과 같은 확률질량함수 $f_\delta(x)$를 구할 수 있다.

$$f_\delta(x : \delta) = P(X = x) = \begin{cases} \delta_+^2 + (1 - \delta^2) f_0(0), & \text{if } x = 0, \\ \left(1 - \delta_+^2 + \delta_-^2 \left\{ \dfrac{f_0(0)}{1 - f_0(0)} \right\}\right) f_0(x), & \text{if } x = 1, 2, \ldots. \end{cases} \tag{2.1}$$

이 때, $\delta_+ = \max(\delta, 0)$, $\delta_- = \max(-\delta, 0)$를 의미한다.

Ghosh와 Kim (2007)에 의하면 확률변수 X의 확률질량함수가 식 (2.1)일 때, 모형은 유일한 표현식을 가지며 산포형태모수 δ를 조정함으로써 과대산포와 과소산포 두 경우 모두에 적용 가능하다.

$x \neq 0$이고 $w(\delta) = 1 - \delta_+^2 + \delta_-^2 \, f_0(0)/(1 - f_0(0))$일 때, $f_\delta(x) = w(\delta) f_0(x)$로 쓸 수 있으며, 확률변수 U의 평균과 분산을 각각 μ_0, σ_0^2이라 할 때, 다음의 결과를 얻게 된다.

$$\begin{aligned} \mu = E(X) &= \sum_{x=0}^{\infty} x f_\delta(x) \\ &= w(\delta)\mu_0 \end{aligned} \tag{2.2}$$

$$\begin{aligned} \sigma^2 = V(X) &= \sum_{x=0}^{\infty} (x - E(X))^2 f_\delta(x) \\ &= w(\delta)\sigma_0^2 + \frac{1 - w(\delta)}{w(\delta)} \mu^2 \end{aligned} \tag{2.3}$$

확률질량함수가 $f_0(x)$인 확률변수 U에 대해 평균 μ_0과 분산 σ_0^2이 같은 균등산포(equi-dispersion)를 가정하고 식 (2.1)을 확률질량함수로 갖는 확률변수 X의 평균과 분산을 각각 μ와 σ^2라 할 때, 산포형태모수 δ에 따른 μ와 σ^2의 관계를 확인하기 위해 σ^2에서 μ를 빼면 다음과 같이 정리할 수 있다.

$$\sigma^2 - \mu = \left\{ w(\delta)\sigma_0^2 + \frac{1 - w(\delta)}{w(\delta)} \mu^2 \right\} - w(\delta)\mu_0 = \frac{1 - w(\delta)}{w(\delta)} \mu^2. \tag{2.4}$$

μ^2은 항상 0보다 크거나 같으므로, $(\sigma^2-\mu)$의 부호는 $\{1-w(\delta)\}/w(\delta)$의 부호와 일치하며

$$\frac{1-w(\delta)}{w(\delta)}=\begin{cases}\dfrac{\delta^2}{1-\delta^2}, & \text{if } \delta>0,\\ 0, & \text{if } \delta=0,\\ \dfrac{-\delta^2 f_0(0)}{1-(1-\delta^2)f_0(0)}, & \text{if } \delta<0\end{cases}$$

로 나타낼 수 있다.

즉, $(\sigma^2-\mu)$와 δ의 부호는 항상 일치하며

i) $\delta<0$이면 $\mu>\sigma^2$으로 과소산포(under-dispersion)

ii) $\delta=0$이면 $\mu=\sigma^2$으로 균등산포(equi-dispersion)

iii) $\delta>0$이면 $\mu<\sigma^2$으로 과대산포(over-dispersion)

임을 확인할 수 있다 (Ghosh와 Kim, 2007).

영 변환 모형의 $x=0$에서의 확률을 $g(0)$라 할 때, $f_\delta(x)$를 확률질량함수로 갖는 분포의 평균과 분산을 각각 μ와 σ^2라 하면 δ의 값은 다음과 같이 표현된다.

$$\delta=\begin{cases}\sqrt{\dfrac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}}, & \text{if } \sigma^2\geq\mu,\\ -\sqrt{\dfrac{\mu-\sigma^2}{\mu-\sigma^2+\frac{g(0)}{1-g(0)}\mu^2}}, & \text{if } \sigma^2<\mu.\end{cases}\tag{2.5}$$

표본에서 δ의 추정치 $\hat{\delta}$를 구하기 위해 위 식에 $\mu,\sigma^2,g(0)$ 대신 표본평균 $\bar{X}$, 표본분산 S^2, 표본에서 0이 차지하는 비율 P_0를 대입하면,

$$\hat{\delta}=\begin{cases}\sqrt{\dfrac{\frac{n-1}{n}S^2-\bar{X}}{\frac{n-1}{n}S^2-\bar{X}+\bar{X}^2}}, & \text{if } \dfrac{n-1}{n}S^2\geq\bar{X},\\ -\sqrt{\dfrac{\bar{X}-\frac{n-1}{n}S^2}{\bar{X}-\frac{n-1}{n}S^2+\frac{P_0}{1-P_0}\bar{X}^2}}, & \text{if } \dfrac{n-1}{n}S^2<\bar{X}.\end{cases}\tag{2.6}$$

여기서 $\bar{X}=\sum_{i=1}^n X_i/n$, $S^2=\sum_{i=1}^n(X-\bar{X})^2/(n-1)$, $P_0=\sum_{i=1}^n I_0(X_i)/n$이 된다.

2.2. 영향함수(influence function)

영향함수는 분포함수 F에 대해 실수값을 갖는 범함수(real-valued functional)를 T라 할 때, 하나의 관찰치가 $T(F)$에 대하여 어느 정도의 영향을 갖는지를 측정하는 함수이다. Hampel (1974)이 영향함수 이론을 소개하였을 때보다 컴퓨터의 성능이 매우 좋아져 관찰치가 제거되었을 때 통계량에 미치는 영향을 예측하는 영향함수의 역할이 최근 크게 주목을 받지는 못하는 상황이지만 학술적인 의의는 크다 할 수 있다.

분포함수 F에 임의의 관찰값 x를 추가하면 다음과 같이 섭동(perturbation)된 분포함수가 된다.

$$F_\epsilon=(1-\epsilon)F+\epsilon\delta_x,\quad 0<\epsilon<1.\tag{2.7}$$

이 때, δ_x는 실수 공간의 한 점 x에서만 확률 1을 갖는 퇴화분포(degenerate distribution) 함수이다.

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

따라서 식 (2.7)과 같이 섭동된 분포함수 F_ϵ에서의 범함수 $T(F)$는 섭동된 범함수 $T(F_\epsilon)$로 정의될 수 있으며, Hampel (1974)은 범함수 $T(F)$에 대한 관찰치 x의 영향함수를 다음과 같이 정의하였다.

$$\begin{aligned} \mathrm{IF}(T, x) &= \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T(F)}{\epsilon}, \end{aligned} \tag{2.8}$$

여기서 $\mathrm{IF}(T, x)$는 $T(F)$의 일차미분계수로 추가된 x에 의해 섭동된 범함수 $T(F_\epsilon)$의 원래 범함수 $T(F)$로부터의 순간변화율을 의미한다. 즉, 식 (2.8)에서 정의한 영향함수 $\mathrm{IF}(T, x)$는 $T(F)$에 대한 x의 고유한 영향을 보여주며, l' Hôspital정리에 의해 다음과 같이 표현할 수 있다.

$$\mathrm{IF}(T, x) = \left[\frac{\partial T(F_\epsilon)}{\partial \epsilon} \right]_{\epsilon=0}. \tag{2.9}$$

모집단의 평균과 분산은 범함수 $T(F)$의 일종이므로 각각을 범함수 $\mu(F)$, $\sigma^2(F)$로 표현할 수 있다.

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

식 (2.8) 또는 식 (2.9)를 이용하여 다음과 같이 평균과 분산에 대한 영향함수를 유도할 수 있다 (Hampel, 1974).

$$\mathrm{IF}(\mu, x) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = x - \mu, \tag{2.10}$$

$$\mathrm{IF}(\sigma^2, x) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = (x-\mu)^2 - \sigma^2 = [\mathrm{IF}(\mu, x)]^2 - \sigma^2. \tag{2.11}$$

한편, 모수 θ^2에 대한 영향함수가 알려져 있을 때 모수 θ (> 0)에 대한 영향함수는 $\theta(F) = \sqrt{\theta^2(F)}$의 관계를 이용하면 다음과 같이 얻어지게 된다.

$$\begin{aligned} \mathrm{IF}(\theta, x) &= \left(\frac{\partial \sqrt{\theta^2(F_\epsilon)}}{\partial \epsilon} \right)_{\epsilon=0} \\ &= \left(\frac{1}{2\sqrt{\theta^2(F_\epsilon)}} \frac{\partial \theta^2(F_\epsilon)}{\partial \epsilon} \right)_{\epsilon=0} \\ &= \frac{1}{2\theta(F)} \mathrm{IF}(\theta^2, x) \\ &= \frac{1}{2\theta} \mathrm{IF}(\theta^2, x). \end{aligned} \tag{2.12}$$

모수 θ_3가 두 모수 θ_1과 θ_2의 함수이고 θ_1과 θ_2에 대한 영향함수가 알려져 있다고 하자. 즉,

$$\theta_3 = g(\theta_1, \theta_2)$$

$$\mathrm{IF}(\theta_1, x) = \left[\frac{\partial\theta_1(F_\epsilon)}{\partial\epsilon}\right]_{\epsilon=0}, \qquad \mathrm{IF}(\theta_2, x) = \left[\frac{\partial\theta_2(F_\epsilon)}{\partial\epsilon}\right]_{\epsilon=0}$$

이다.

이제 θ_3에 대한 영향함수는

$$\mathrm{IF}(\theta_3, x) = \left[\frac{\partial\theta_3(F_\epsilon)}{\partial\epsilon}\right]_{\epsilon=0}$$

로 주어지며 두 변수 함수의 미분에 관한 연쇄법칙(chain rule)에 의해

$$\begin{aligned}\frac{\partial\theta_3(F_\epsilon)}{\partial\epsilon} &= \frac{\partial g(\theta_1(F_\epsilon), \theta_2(F_\epsilon))}{\partial\epsilon} \\ &= \frac{\partial g}{\partial\theta_1}\cdot\frac{\partial\theta_1(F_\epsilon)}{\partial\epsilon} + \frac{\partial g}{\partial\theta_2}\cdot\frac{\partial\theta_2(F_\epsilon)}{\partial\epsilon}\end{aligned}$$

이 되고 결과적으로

$$\mathrm{IF}(\theta_3, x) = \frac{\partial\theta_3}{\partial\theta_1}\mathrm{IF}(\theta_1, x) + \frac{\partial\theta_3}{\partial\theta_2}\mathrm{IF}(\theta_2, x) \tag{2.13}$$

를 얻게된다.

3. 영 변환 모형에서의 영향함수

3.1. 영 변환 모형 모수 δ에 대한 영향함수 유도

두 모수 μ와 σ^2으로 구성되어 있는 δ^2의 영향함수 $\mathrm{IF}(\delta^2, x)$는 식 (2.13)에서 주어진 결과를 이용하면 다음과 같이 μ와 σ^2에 대한 영향함수의 합으로 나타낼 수 있으며,

$$\begin{aligned}\mathrm{IF}\left(\delta^2, x\right) &= \frac{\partial\delta^2}{\partial\mu}\mathrm{IF}(\mu, x) + \frac{\partial\delta^2}{\partial\sigma^2}\mathrm{IF}\left(\sigma^2, x\right) \\ &= \frac{\mu^2 - 2\mu\sigma^2}{\left(\sigma^2 - \mu + \mu^2\right)^2}(x-\mu) + \frac{\mu^2}{\left(\sigma^2 - \mu + \mu^2\right)^2}\left\{(x-\mu)^2 - \sigma^2\right\}.\end{aligned} \tag{3.1}$$

식 (2.12)로 주어진 결과인 $\mathrm{IF}(\theta, x) = 1/(2\theta(F))\,\mathrm{IF}(\theta^2, x)$에 식 (3.1)을 대입하면, δ의 영향함수 식 (3.2)가 얻어진다.

$$\begin{aligned}\mathrm{IF}(\delta, x) &= \frac{1}{2\delta}\mathrm{IF}\left(\delta^2, x\right) \\ &= \frac{1}{2\sqrt{\frac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}}}\left[\frac{\mu^2 - 2\mu\sigma^2}{\left(\sigma^2 - \mu + \mu^2\right)^2}(x-\mu) + \frac{\mu^2}{\left(\sigma^2 - \mu + \mu^2\right)^2}\left\{(x-\mu)^2 - \sigma^2\right\}\right].\end{aligned} \tag{3.2}$$

3.2. 실제 자료 적용

이 절에서는 인구주택총조사 자료 중 '연령 및 출생자녀수별 기혼여성인구(前 시도/연령/출생아수별 부인)' 자료를 이용하여 Ghosh와 Kim의 산포형태모수 δ를 추정하고, 특정 데이터를 제거했을 경우 δ의 실제 변화량과 앞에서 유도한 영 변환 모형의 영향함수로 구한 추정값의 비교를 통해 유도한 영향함수가 적절한지 알아보았다. Kim 등 (2013)의 결과를 보면 δ값이 $+$와 $-$값을 경향성을 가지고 변하고 있으며 이는 자료가 영 변환 모형에 적합하다는 증거가 된다.

Table 3.1. The number of babies born to married women aged 65 to 69(1975)

출생아수별 부인수(미상제외)											출생아수		$\hat{\delta}$
0명	1명	2명	3명	4명	5명	6명	7명	8명	9명	10명	$\bar{X}$	S^2	
12,604	20,918	23,604	34,800	45,292	47,502	44,483	32,583	26,040	14,813	15,074	4.969	6.514	0.243

'연령 및 출생자녀수별 기혼여성인구(15세 이상)' 자료는 1970년부터 5년마다 각 연도별로 5–10% 표본에서 조사한 출산력/여성·아동 부문의 표본조사자료(1995년은 미실시)로 조사대상은 15세 이상의 기혼여성이며, 출생아수 구분이 다른 1970년을 제외하고는 15세부터 75세 이상 까지 5세 단위로 구분하고 출생아수는 0명에서 10명 이상까지 11개의 그룹으로 분류되어 있다. 본 논문에서는 이 중 평균과 분산의 차이가 가장 큰 1975년 65–69세 데이터를 $\sigma^2 \geq \mu$인 영 변환 모형 영향함수의 타당성을 검정하기 위한 자료로 선택하고 계산상 편의를 위해 10명 이상의 출생아수는 10명에 포함하였다.

Ghosh와 Kim 모수 δ의 영향함수에 대한 타당성을 검증하고자 위 자료의 원데이터와 앞 절에서 구한 $\bar{X}$, S^2, $\hat{\delta}$을 정리하면 Table 3.1과 같다.

출생아수 0명, 1명, . . . , 10명 중 어느 값에서 데이터가 제거되었을 때 $\hat{\delta}$에 가장 큰 변화를 주는지를 확인하기 위해 각각의 경우에서 기혼여성 수 중 출생아수별 기혼여성 수를 1명, 10명, 100명, 1,000명, 5,000명, 10,000명 제거했을 때의 $\hat{\delta}$과 영향함수 IF(δ, x) 값을 구하고 제거하는 데이터 수에 따른 값의 변화 정도를 살펴보았다. 이 때 제거한 데이터 수에 따른 가중치를 영향함수에 반영하기 위해 IF(δ, x)에 $n/(N-1)$을 곱한 값을 $\hat{\delta}$ 변화에 대한 추정치로 사용하게 된다.

$\hat{\delta}$와 $\hat{\delta}_{i-}$의 차이를 보면 대부분의 경우에서 출생아수가 극단적인 값, 0명 또는 10명에 가까울수록 차이가 크게 나타났음을 확인할 수 있다 (Table 3.2). 또한 데이터를 1개씩 제거했을 경우 $\hat{\delta}$의 최대 변화율 $|(1 - \hat{\delta}_{i-}/\hat{\delta}) * 100|$은 출생아수가 0명인 자료를 제거했을 때는 0.0023664%밖에 되지 않아 이 자료에서 데이터 1개가 가진 영향력은 매우 작다는 것을 확인할 수 있었다. 반면 출생아수별로 데이터를 10,000개씩 제거했을 때에는 최대 변화율이 27.78%나 되었다. 정리하면 출생아수별 기혼여성 수 자료에서는 제거하는 데이터가 자료 분포에서 양 끝에 위치하고 제거하는 수가 클수록 $\hat{\delta}$의 변화에 큰 영향을 미친다고 할 수 있다.

Ghosh와 Kim의 산포형태모수 $\hat{\delta}$의 변화량을 계산한 두 값, 실제값$(\hat{\delta} - \hat{\delta}_{i-})$과 추정값(IF$(\delta, x)$)이 거의 일치하므로 (Table 3.2), 앞 장에서 유도한 영향함수가 타당한 것으로 보이는데, 보다 정확한 검증을 위해 실제값$(\hat{\delta} - \hat{\delta}_{i-})$과 추정값(IF$(\delta, x)$)에 대한 회귀분석을 실시한 결과, 모든 경우에서 $\hat{\delta}$의 변화량을 실제로 계산한 값과 영향함수를 이용해 추정한 값이 유의수준 0.05에서 선형관계를 가지며, 1개의 데이터를 제거했을 때에는 회귀계수가 1이고 R^2이 100%으로 영향함수가 δ의 변화를 거의 완벽하게 측정하였다고 볼 수 있다.

제거한 데이터 수가 10개, 100개, . . . , 10,000개로 늘어날수록 회귀계수값과 R^2가 점점 줄어드는 것을 볼 수 있는데 (Table 3.3) 이를 통해 데이터 감소가 클수록 영향함수의 정확도가 감소함을 확인할 수 있었다. 따라서 영 변환 모형의 산포형태모수 δ에 대하여 3.1절에서 유도한 영향함수는 제거하는 데이터 수의 영향을 받긴 하지만 추정된 변화량과 실제 계산된 변화량이 매우 근사한 값을 가지므로 δ에 대한 데이터의 영향력을 추정하는데에 있어 매우 타당한 방법으로 판단된다.

4. 결론

본 논문에서는 영 과잉과 영 부족의 경우를 모두 포함하여 설명할 수 있는 Ghosh와 Kim 영 변환(zero-altered) 모형의 산포형태모수(dispersion type parameter) δ에 대한 영향함수 IF(δ, x)를 다음과 같이

Table 3.2. Comparison of Ghosh and Kim's parameters $\hat{\delta}$ and IF(δ, x) according to the number of removed data

제거한 데이터수	출생아수 제거구간	$\hat{\delta}_{i-}$	실제값 $(\hat{\delta} - \hat{\delta}_{i-})$	추정값 (IF(δ, x))	실제값 / 추정값	실제값 − 추정값
원본 $\hat{\delta}$		0.2426941				
	0	0.2426883	0.0000057	0.0000061	0.9410951	−0.0000004
	1	0.2426908	0.0000033	0.0000036	0.9014161	−0.0000004
	2	0.2426928	0.0000013	0.0000017	0.7827964	−0.0000004
	3	0.2426943	−0.0000002	0.0000001	−1.7877482	−0.0000004
	4	0.2426954	−0.0000013	−0.0000009	1.3858215	−0.0000004
1	5	0.2426960	−0.0000019	−0.0000015	1.2354102	−0.0000004
	6	0.2426961	−0.0000020	−0.0000017	1.2169404	−0.0000004
	7	0.2426958	−0.0000017	−0.0000013	1.2719678	−0.0000004
	8	0.2426950	−0.0000009	−0.0000005	1.6897049	−0.0000004
	9	0.2426937	0.0000004	0.0000007	0.5172389	−0.0000004
	10	0.2426920	0.0000021	0.0000025	0.8547887	−0.0000004
	0	0.2426366	0.0000574	0.0000610	0.9409948	−0.0000036
	1	0.2426612	0.0000329	0.0000365	0.9013588	−0.0000036
	2	0.2426811	0.0000130	0.0000165	0.7827534	−0.0000036
	3	0.2426964	−0.0000023	0.0000013	−1.7899596	−0.0000036
	4	0.2427070	−0.0000129	−0.0000093	1.3859100	−0.0000036
10	5	0.2427129	−0.0000189	−0.0000153	1.2355220	−0.0000036
	6	0.2427142	−0.0000202	−0.0000166	1.2170866	−0.0000036
	7	0.2427109	−0.0000168	−0.0000132	1.2722130	−0.0000036
	8	0.2427029	−0.0000088	−0.0000052	1.6908068	−0.0000036
	9	0.2426902	0.0000039	0.0000075	0.5168320	−0.0000036
	10	0.2426729	0.0000212	0.0000248	0.8544593	−0.0000036
	0	0.2421189	0.0005752	0.0006119	0.9399897	−0.0000367
	1	0.2423652	0.0003289	0.0003651	0.9007856	−0.0000362
	2	0.2425646	0.0001295	0.0001655	0.7823232	−0.0000360
	3	0.2427173	−0.0000232	0.0000128	−1.8122323	−0.0000360
	4	0.2428234	−0.0001293	−0.0000932	1.3867947	−0.0000361
100	5	0.2428829	−0.0001888	−0.0001527	1.2366402	−0.0000361
	6	0.2428959	−0.0002019	−0.0001656	1.2185485	−0.0000362
	7	0.2428624	−0.0001683	−0.0001321	1.2746669	−0.0000363
	8	0.2427823	−0.0000882	−0.0000518	1.7018928	−0.0000364
	9	0.2426555	0.0000385	0.0000752	0.5127956	−0.0000366
	10	0.2424820	0.0002121	0.0002491	0.8511765	−0.0000371
	0	0.2368639	0.0058302	0.0062713	0.9296650	−0.0004411
	1	0.2393816	0.0033125	0.0037012	0.8949745	−0.0003887
	2	0.2414025	0.0012915	0.0016600	0.7780171	−0.0003685
	3	0.2429402	−0.0002461	0.0001199	−2.0519994	−0.0003660
	4	0.2440040	−0.0013100	−0.0009386	1.3956228	−0.0003713
1,000	5	0.2446001	−0.0019061	−0.0015275	1.2478203	−0.0003785
	6	0.2447309	−0.0020369	−0.0016517	1.2331952	−0.0003852
	7	0.2443956	−0.0017016	−0.0013095	1.2994358	−0.0003921
	8	0.2435900	−0.0008959	−0.0004923	1.8198364	−0.0004036
	9	0.2423064	0.0003877	0.0008153	0.4754952	−0.0004276
	10	0.2405334	0.0021607	0.0026368	0.8194384	−0.0004761

(continued)

(Continued)

제거한 데이터수	출생아수 제거구간	$\hat{\delta}_{i-}$	실제값 $(\hat{\delta}-\hat{\delta}_{i-})$	추정값 $(\mathrm{IF}(\delta,x))$	실제값 / 추정값	실제값 − 추정값
5,000	0	0.2115483	0.0311457	0.0355310	0.8765794	−0.0043853
	1	0.2255765	0.0171176	0.0197382	0.8672311	−0.0026206
	2	0.2363186	0.0063755	0.0084022	0.7587902	−0.0020267
	3	0.2442378	−0.0015437	0.0004171	−3.7007972	−0.0019608
	4	0.2496146	−0.0069206	−0.0048245	1.4344715	−0.0020961
	5	0.2526064	−0.0099123	−0.0076400	1.2974212	−0.0022723
	6	0.2532753	−0.0105812	−0.0081469	1.2987964	−0.0024343
	7	0.2515998	−0.0089058	−0.0062950	1.4147269	−0.0026107
	8	0.2474723	−0.0047782	−0.0018590	2.5703619	−0.0029193
	9	0.2406837	0.0020104	0.0056104	0.3583317	−0.0036000
	10	0.2308887	0.0118054	0.0169022	0.6984524	−0.0050968
10,000	0	0.1737041	0.0689900	0.0878646	0.7851852	−0.0188746
	1	0.2068592	0.0358349	0.0433218	0.8271806	−0.0074868
	2	0.2301700	0.0125241	0.0170492	0.7345858	−0.0045251
	3	0.2465694	−0.0038754	0.0003720	−10.4189188	−0.0042473
	4	0.2574324	−0.0147383	−0.0099438	1.4821637	−0.0047945
	5	0.2634275	−0.0207334	−0.0152570	1.3589478	−0.0054765
	6	0.2647994	−0.0221053	−0.0160006	1.3815316	−0.0061047
	7	0.2614583	−0.0187642	−0.0119401	1.5715244	−0.0068241
	8	0.2529619	−0.0102679	−0.0021031	4.8823539	−0.0081648
	9	0.2383819	0.0043121	0.0156737	0.2751179	−0.0113616
	10	0.2159331	0.0267609	0.0461530	0.5798310	−0.0193921

Table 3.3. Regression analysis of actual value and estimated value according to the number of removed data

n	1	10	100	1,000	5,000	10,000
회귀계수	1.0000	0.9999	0.9993	0.9925	0.9511	0.8672
R^2	100.0%	100.0%	100.0%	100.0%	99.7%	98.7%

유도하였다.

$$\begin{aligned}\mathrm{IF}(\delta,x) &= \frac{1}{2\delta}\mathrm{IF}\left(\delta^2,x\right) = \frac{1}{2\sqrt{\frac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}}}\mathrm{IF}\left(\delta^2,x\right)\\ &= \frac{1}{2\sqrt{\frac{\sigma^2-\mu}{\sigma^2-\mu+\mu^2}}}\left[\frac{\mu^2-2\mu\sigma^2}{\left(\sigma^2-\mu+\mu^2\right)^2}(x-\mu)+\frac{\mu^2}{\left(\sigma^2-\mu+\mu^2\right)^2}\left\{(x-\mu)^2-\sigma^2\right\}\right].\end{aligned}$$

유도된 δ에 대한 영향함수의 타당성 검증을 위해 '우리나라 기혼여성의 출생아수' 자료를 가지고 산포형 태모수 추정치 $\hat{\delta}$의 변화를 직접 구하고 영향함수로 계산한 값과 비교해보았다. 그 결과, 두 값이 유의수준 0.05에서 일치하는 것으로 나타나 유도된 영향함수가 타당함을 확인하였다. 이러한 결과를 이용하면 향후 Ghosh와 Kim의 영 변환 모형의 산포형태모수를 추정하는 경우 미리 이 모수의 추정에 큰 영향을 미치는 자료값을 알아봄으로써 정확한 모수 추정에 좀 더 다가갈 수 있으리라 기대할 수 있다.

References

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

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

Cook, R. D. and Weisberg, S. (1982). *Residual and Influence in Regression*, Chapman and Hall, New York.

Critchley, F. (1985). Influence in principal components analysis, *Biometrika*, **72**, 627–626.

Ghosh, S. K. and Kim, H. (2007). Semiparametric inference based on a class of zero-altered distributions, *Statistical Methodology*, **4**, 371–383.

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

Kim, H. (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, **40**, 201–207.

Kim, H. (1994). Influence functions in multiple correspondence analysis, *The Korean Journal of Applied Statistics*, **7**, 69–74.

Kim, H. (1998). A study on cell influences to χ^2 statistics in contingency tables, *The Korean Communications in Statistics*, **5**, 35–42.

Kim, H. and Lee, H. (1996). Influence function on χ^2 statistics in contingency tables, *The Korean Communications in Statistics*, **7**, 69–76.

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

Kim, H., Ra, Y., Kim, K., and Lee, Y. (2013). Analysis of the trend of the number of children of married women in Korea using zero-altered distribution, *Journal of Korean Official Statistics*, **18**, 1–15.

Kim, K. (2005). The influence function in t statistic and the test of its validity (master's thesis), Chungnam National University, Daejeon.

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

Ghosh와 Kim 모수 δ의 영향함수 유도 및 확인

김민정[a] · 김홍기[a,1]

[a]통계정 소득통계과, [b]충남대학교 정보통계학과

(2017년 4월 10일 접수, 2017년 6월 26일 수정, 2017년 7월 27일 채택)

요 약

Ghosh와 Kim에 의해 소개된 영 변환 모형은 0이 많거나 적을 때 계수형 자료(count data)를 분석하는 모형이다. 이 모형의 산포형태모수는 평균과 분산, 0 확률로 구성되며 μ와 σ^2의 관계에 따라 2가지 형태를 가진다. 본 논문에서는 $\sigma^2 \geq \mu$일 때, Ghosh와 Kim 영 변환확률 모형의 모수 δ에 대한 영향함수를 도출하였다. 도출한 영향함수의 타당성을 검증하기 위해서 인구주택총조사 자료를 이용해 관측치가 제거된 경우에서 영향함수로 도출한 δ 추정치 변화값과 직접 계산한 δ 추정치 변화값을 비교하였다. 그 결과 영향함수는 δ의 변화를 매우 정확히 추정하였다.

주요용어: 과대산포, 과소산포, 산포형태모수, 영 변환 모형, 영향함수

이 연구는 2016년도 충남대학교 학술 연구비에 의해 지원되었음.

[1]교신저자: (34134) 대전시 유성구 대학로 99, 충남대학교 정보통계학과. E-mail: honggiekim@cnu.ac.kr