The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.5.527
33(5), 527–540

A study on the difference and calibration of empirical influence function and sample influence function

Hyunseok Kang[a] · Honggie Kim[b,1]

[a]Daejeon High School; [b]Department of Information and Statistics, Chungnam National University

(Received May 20, 2020; Revised June 18, 2020; Accepted July 3, 2020)

Abstract

While analyzing data, researching outliers, which are out of the main tendency, is as important as researching data that follow the general tendency. In this study we discuss the influence function for outlier discrimination. We derive sample influence functions of sample mean, sample variance, and sample standard deviation, which were not directly derived in previous research. The results enable us to mathematically examine the relationship between the empirical influence function and sample influence function. We can also consider a method to approximate the sample influence function by the empirical influence function. Also, the validity of the relationship between the approximated sample influence function and the empirical influence function is also verified by the simulation of random sampled data in normal distribution. As the result of a simulation, both the relationship between the two influence functions, sample and empirical, and the method of approximating the sample influence function through the emperical influence function were verified. This research has significance in proposing a method that reduces errors in the approximation of the empirical influence function and in proposing an effective and practical method that proceeds from previous research that approximates the sample influence function directly through empirical influence function by constant revision.

Keywords: influence function, outlier, empirical influence function, sample influence function

1. 서론

데이터들이 이루는 보편적인 경향과 분포와는 달리 그 경향성과 분포를 벗어나 특별히 크거나 작은 값을 갖는 데이터들을 우리는 이상치(outlier)라 부른다. 이상치에 대한 적절한 선별과 배제 없이 모든 데이터를 종합적으로 분석하게 되는 경우 데이터 분석을 통해 얻은 결과의 신뢰성과 해석의 일반성에 치명적인 위협을 받을 수 있다. 따라서 데이터의 분석 과정에서 이러한 이상치를 판별하고, 이상치가 통계량, 통계적 모형에 어떠한 영향을 주는 지에 대한 분석은 매우 중요한 일이라 할 수 있다. Hampel (1974)은 영향함수(influence function)를 활용하여 이상치를 판별할 수 있는 방법을 가장 먼저 소개하였다. 영향함수는 전체의 데이터에서 하나의 관측값이 제외되거나 추가됨으로써 관심 있는 통계량에 어느 정도의 영향을 미치는지 그 변화에 대한 상대적인 척도를 나타내는 함수로서 전체 데이터에 미치는 영향의 정도를 쉽게 알 수 있는 함수이다. 특히 이 영향함수는 이상치를 판별하는데 주로 사용되고 있는 함수

[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

로 Hampel은 영향함수를 통해 거의 모든 통계량에서 이상치를 판별할 수 있음을 보였다. 이후, Campbell (1978)은 판별분석에서 영향함수를 사용하여 이상치를 발견하였고, Radhakrishnan과 Kshirsagar (1981)은 다변량 분석에서 여러 가지 모수에 대해 이론적으로 영향함수를 유도해 냈다. Cook (1977)은 회귀 분석에서의 영향력 있는 관측값에 대해 연구하였으며, Cook과 Weisberg (1980, 1982)는 회귀분석에서 회귀진단방법으로서 영향함수를 적용하였다. Critchley (1985)는 주성분분석에서 영향함수를 적용하여 영향력 있는 관측치를 찾아내는 방법으로 활용하였다. Kim과 Lee (1996), Kim (1998), Lee와 Kim (2003) 등은 χ^2 통계량에 대한 영향함수, Kim과 Kim (2005)은 t 통계량에 대한 영향함수, Lee와 Kim (2008)의 변이계수에 대한 영향함수의 유도까지 다양한 통계량에 대한 영향함수를 유도해 내는 연구가 활발히 이루어져 왔다. 최근에는 Kim과 Kim (2019)의 빅데이터에서 모분포의 형태에 따른 t 통계량에 대한 영향함수의 성능에 대한 연구와 Park와 Kim (2019)의 t 통계량에 가장 작은 영향을 미치는 관측값의 위치에 대한 연구에 이르기까지 영향함수의 유도와 다양한 분야에서의 활용에 대한 연구가 지금까지도 계속 활발하게 진행되고 있다.

본 연구에서는 표본평균, 표본분산, 표본표준편차에 대한 표본영향함수를 직접 유도하고, 경험적 영향함수와의 차이를 확인하여 이상치를 판별하기 위한 방법론으로서의 영향함수 활용 과정에 엄밀성을 높여 보고자 한다. 표본평균, 표본분산, 표본표준편차는 t 통계량과 함께 사회과학 연구에서 그 활용도가 매우 높기 때문에 본 연구는 각 관측값들이 표본평균, 표본분산, 표본표준편차에 미치는 영향을 엄밀하게 확인함으로써 이상치를 제거해야할 경우, 이상치 제거를 위한 우선순위를 정할 수 있도록 도울 수 있다. 2장에서는 영향함수의 정의와 평균, 분산, 표준편차에 대한 영향함수와 경험적 영향함수의 유도와 함께 표본영향함수를 정의한다. 3장에서는 표본평균, 표본분산, 표본표준편차에 대한 경험적 영향함수와 표본영향함수를 각각 유도하고 이에 대한 관계를 이론적으로 살펴본다. 4장에서는 모의로 생성한 데이터를 기반으로 3장에서 유도한 경험적 영향함수와 표본영향함수 추론의 타당성을 검증한 뒤, 5장에서는 실제 자료 분석 과정에 이를 적용한 예를 다룬다. 6장에서는 본 연구의 결론을 제시한다.

2. 영향함수

2.1. 영향함수의 정의 및 평균, 분산, 표준편차의 영향함수

분포함수 F에 대해 $T(F) = c$와 같이 실숫값 c를 함숫값으로 갖는 함수 T를 범함수(real-valued function)라 하고, 실수 공간의 한 점인 x에서 확률이 1인 분포함수

$$\delta_x(t) = \begin{cases} 0, & t < x, \\ 1, & t \geq x \end{cases} \tag{2.1}$$

를 퇴화분포함수(degenerated distribution function)라고 한다. $F(t)$를 F로, $\delta_x(t)$를 δ_x로, $F_\epsilon(t)$를 F_ϵ으로 각각 표기하면 분포함수 F에 임의의 관측값 x를 추가함으로써 생기는 분포함수 F와 퇴화분포함수 δ_x의 혼합분포함수 F_ϵ는 다음과 같다.

$$F_\epsilon = (1-\epsilon)F + \epsilon\delta_x, \quad 0 < \epsilon < 1. \tag{2.2}$$

이때, F_ϵ를 F의 섭동(perturbation)이라고 한다.

Hampel (1974)은 관측값 x가 추가됨으로써 범함수 $T(F)$에 미치는 영향을 나타내는 영향함수 $\mathrm{IF}(T, x)$를 분포함수 F의 섭동 F_ϵ를 이용해 다음과 같이 정의하였다.

$$\mathrm{IF}(T, x) = \lim_{\epsilon \to 0} \frac{T\left(F_\epsilon\right) - T(F)}{\epsilon} = \lim_{\epsilon \to 0} \frac{T[(1-\epsilon)F + \epsilon\delta_x] - T(F)}{\epsilon}. \tag{2.3}$$

식 (2.3)을 살펴보면 영향함수 $\mathrm{IF}(T,x)$는 분포 F에 대하여 관측값 x에서의 섭동된 범함수 $T(F_\epsilon)$에 의한 $T(F)$의 일차미분계수 형태, 즉 순간변화율을 나타낸다. 로피탈의 정리를 이용해 식 (2.3)을 계산하면 다음과 같다.

$$\mathrm{IF}(T,x)=\lim_{\epsilon\to 0}\frac{T(F_\epsilon)-T(F)}{\epsilon}=\lim_{\epsilon\to 0}\left[\frac{\partial T(F_\epsilon)}{\partial\epsilon}\right]=\left[\frac{\partial T(F_\epsilon)}{\partial\epsilon}\right]_{\epsilon=0}. \tag{2.4}$$

한편, 모집단의 분포가 갖는 평균 μ와 분산 σ^2에 대한 범함수를 각각 T_1, T_2라 하고, 모집단의 분포함수 $F(t)$의 확률밀도함수를 $f(t)$라 하면 $\partial F(t)/\partial t=f(t)$가 성립하므로 모집단의 평균과 분산을 다음과 같이 범함수의 형태로 나타낼 수 있다.

$$\begin{aligned}T_1(F)&=\mu=\int tf(t)dt=\int tdF(t),\\ T_2(F)&=\sigma^2=\int(t-\mu)^2f(t)dt=\int(t-\mu)^2dF(t).\end{aligned} \tag{2.5}$$

식 (2.4)와 식 (2.5)를 이용해 평균 μ와 분산 σ^2에 대한 영향함수를 유도하면 다음과 같다 (Hampel, 1974).

$$\begin{aligned}\mathrm{IF}(T_1,x)&=\mathrm{IF}(\mu,x)=x-\mu,\\ \mathrm{IF}(T_2,x)&=F\left(\sigma^2,x\right)=(x-\mu)^2-\sigma^2.\end{aligned} \tag{2.6}$$

$T_3(F)=\sqrt{T_2(F)}=\sigma$는 영향함수 $\mathrm{IF}(T_2,x)$를 이용하면 다음과 같이 정리할 수 있다.

$$\begin{aligned}\mathrm{IF}(T_3,x)&=\left[\frac{1}{2\sqrt{T_2(F_\epsilon)}}\times\frac{\partial T_2(F_\epsilon)}{\partial\epsilon}\right]_{\epsilon=0}\\ &=\frac{1}{2\sigma}\mathrm{IF}\left(T_2,x\right).\end{aligned} \tag{2.7}$$

즉, 식 (2.7)에 의해 $\mathrm{IF}(T_3,x)=\mathrm{IF}(\sigma,x)=\{1/(2\sigma)\}\cdot\{(x-\mu)^2-\sigma^2\}$로 표준편차 σ에 대한 영향함수를 유도할 수 있다.

2.2. 경험적 영향함수와 표본영향함수

모집단의 분포함수 F에 대한 $T(F)$의 영향함수 $\mathrm{IF}(T,x)$가 정의된다면 표본분포함수 $\hat{F}$에서 얻은 범함수 $T(\hat{F})$의 영향함수는 영향함수 $\mathrm{IF}(T,x)$에 모분포의 통계량을 추정하는 추정량을 대입하여 얻고, 이를 경험적 영향함수(empirical influence function; EIF)라 한다. 모분포가 갖는 통계량인 평균 μ, 표준편차 σ를 추정하는 추정량을 각각 표본평균 $\bar{x}$, 표본표준편차 s라 하면 식 (2.6), 식 (2.7)에 의하여 표본평균, 표본분산, 표본표준편차의 경험적 영향함수는 다음과 같다.

$$\begin{aligned}\mathrm{EIF}(\bar{x},x)&=x-\bar{x},\\ \mathrm{EIF}(s^2,x)&=(x-\bar{x})^2-s^2,\\ \mathrm{EIF}(s,x)&=\frac{1}{2s}\left\{(x-\bar{x})^2-s^2\right\}.\end{aligned} \tag{2.8}$$

표본의 크기가 n이고, 표본평균이 $\bar{x}$, 표본분포함수가 $\hat{F}$인 표본에서 i번째 관측값 x_i를 제거한 표본의 크기 $n-1$인 표본의 표본평균과 표본분포함수를 각각 $\bar{x}_{(i)}$, $\hat{F}_{(i)}$라 하자. 이때, i번째 관측치를 제거함으로써 생기는 범함수의 함숫값의 차이에 섭동인 ϵ을 $-1/(n-1)$로 고려하여 다시 얻어지는 영향함수

를 표본영향함수(sample influence function; SIF)라 한다. Cook과 Weisberg (1982)에 의하면 이는 하나의 관측치가 표본의 통계량에 미치는 영향을 측정하는 도구가 된다. 범함수 T에 대하여 $\mathrm{SIF}(T, x_i)$는 $T(\hat{F}_{(i)}) - T(\hat{F}) = \{-1/(n-1)\} \cdot \mathrm{SIF}(T, x_i)$와 같이 정의한다. 따라서 표본의 크기가 n인 표본에서 i번째 관측값 x_i를 제거한 후의 표본평균, 표본분산, 표본표준편차를 각각 $\bar{x}_{(i)}$, $s^2_{(i)}$, $s_{(i)}$라 하면 범함수에 대한 표본영향함수 $\mathrm{SIF}(T, x_i)$는 다음과 같이 정의할 수 있다.

$$\begin{aligned} \mathrm{SIF}(\bar{x}, x_i) &= -(n-1)\{\bar{x}_{(i)} - \bar{x}\}, \\ \mathrm{SIF}(s^2, x_i) &= -(n-1)\left\{s^2_{(i)} - s^2\right\}, \\ \mathrm{SIF}(s, x_i) &= -(n-1)\{s_{(i)} - s\}. \end{aligned} \tag{2.9}$$

3. 표본영향함수의 유도

표본평균의 표본영향함수를 유도하기 위해 $T(F) = \mu$, $T(\hat{F}) = \bar{x}$로 정의하면 $T(\hat{F}) = (1/n)\sum_{k=1}^{n} x_k$, $T(\hat{F}_{(i)}) = \{1/(n-1)\} \cdot (-x_i + \sum_{k=1}^{n} x_k)$로 나타낼 수 있으므로 다음이 성립한다.

$$\bar{x}_{(i)} - \bar{x} = \frac{1}{n-1}\left(-x_i + \frac{1}{n}\sum_{k=1}^{n} x_k\right) = -\frac{1}{n-1}(x_i - \bar{x}). \tag{3.1}$$

이때, $\mathrm{EIF}(\bar{x}, x_i) = x_i - \bar{x}$이고 $\mathrm{SIF}(\bar{x}, x_i) = -(n-1)\{T(\hat{F}_{(i)}) - T(\hat{F})\} = x_i - \bar{x}$이므로 표본평균 $\bar{x}$에 대한 경험적 영향함수와 표본영향함수는 같다.

다음으로 표본분산, 표본표준편차의 표본영향함수 유도에 앞서 x_i에 대한 함수 $A(x_i)$와 $B(x_i)$를 다음과 같이 정의하자.

$$\begin{aligned} A(x_i) &= -\{x_i - \bar{x}_{(i)}\}^2 + \sum_{k=1}^{n}\left\{x_k - \bar{x}_{(i)}\right\}^2, \\ B(x_i) &= -(x_i - \bar{x})^2 + \sum_{k=1}^{n}(x_k - \bar{x})^2 \end{aligned} \tag{3.2}$$

$\bar{x}_{(i)}$와 $\bar{x}$에 대해서는 $\bar{x}_{(i)} - \bar{x} = \{1/(n-1)\} \cdot (\bar{x} - x_i)$와 $\bar{x}_{(i)} + \bar{x} = \{1/(n-1)\} \cdot \{(2n-1)\bar{x} - x_i\}$이 성립하므로 $A(x_i) - B(x_i)$를 다음과 같이 정리할 수 있다.

$$\begin{aligned} A(x_i) - B(x_i) &= -\{x_i - \bar{x}_{(i)}\}^2 + (x_i - \bar{x})^2 + \sum_{k=1}^{n}\left[\{x_k - \bar{x}_{(i)}\}^2 - (x_k - \bar{x})^2\right] \\ &= \{\bar{x}_{(i)} - \bar{x}\}\left[2x_i - \{\bar{x}_{(i)} + \bar{x}\} - 2\sum_{k=1}^{n} x_k + \sum_{k=1}^{n}\{\bar{x}_{(i)} + \bar{x}\}\right] \\ &= -\frac{1}{n-1}(x_i - \bar{x})^2. \end{aligned} \tag{3.3}$$

즉, $A(x_i) = B(x_i) - \{1/(n-1)\} \cdot (x_i - \bar{x})^2$이 성립하기 때문에 다음의 식이 성립한다.

$$\begin{aligned} -\left\{x_i - \bar{x}_{(i)}\right\}^2 + \sum_{k=1}^{n}\left\{x_k - \bar{x}_{(i)}\right\}^2 &= -(x_i - \bar{x})^2 + \sum_{k=1}^{n}(x_k - \bar{x})^2 - \frac{1}{n-1}(x_i - \bar{x})^2 \\ &= -\frac{n}{n-1}(x_i - \bar{x})^2 + \sum_{k=1}^{n}(x_k - \bar{x})^2. \end{aligned} \tag{3.4}$$

다음으로 표본분산의 표본영향함수를 유도하기 위해 $T(F)=\sigma^2$, $T(\hat{F})=s^2$로 정의하면 각각

$$T(\hat{F})=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2,\quad T(\hat{F}_{(i)})=\frac{1}{n-2}\left[-\{x_i-\bar{x}_{(i)}\}^2+\sum_{k=1}^{n}\{x_k-\bar{x}_{(i)}\}^2\right]$$

로 나타낼 수 있으므로 식 (3.4)를 이용하여 정리하면

$$\begin{aligned}s_{(i)}^2-s^2&=\frac{1}{n-2}\left\{-\frac{n}{n-1}(x_i-\bar{x})^2+\sum_{k=1}^{n}(x_k-\bar{x})^2\right\}-\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2\\&=-\frac{n}{(n-1)(n-2)}\left\{(x_i-\bar{x})^2-s^2\right\}-\frac{s^2}{(n-1)(n-2)}\end{aligned}\tag{3.5}$$

이 성립한다. 이때, 식 (2.8)에 의해 $\mathrm{EIF}(s^2,x_i)=(x_i-\bar{x})^2-s^2$이므로

$$s_{(i)}^2-s^2=-\frac{n}{(n-1)(n-2)}\mathrm{EIF}\left(s^2,x_i\right)-\frac{s^2}{(n-1)(n-2)}\tag{3.6}$$

이고, 표본영향함수의 정의에 의해

$$-\frac{n}{(n-1)(n-2)}\mathrm{EIF}\left(s^2,x_i\right)-\frac{s^2}{(n-1)(n-2)}=-\frac{1}{n-1}\mathrm{SIF}\left(s^2,x_i\right)\tag{3.7}$$

이므로

$$\mathrm{SIF}\left(s^2,x_i\right)=\frac{n}{n-2}\mathrm{EIF}\left(s^2,x_i\right)+\frac{s^2}{n-2}$$

이다. 마지막으로 $T(F)=\sigma$, $T(\hat{F})=s$로 정의하면 각각

$$T\left(\hat{F}\right)=\sqrt{\frac{\sum_{k=1}^{n}\left(x_k-\bar{x}\right)^2}{n-1}},\quad T\left(\hat{F}_{(i)}\right)=\sqrt{\frac{-\left\{x_i-\bar{x}_{(i)}\right\}^2+\sum_{k=1}^{n}\left\{x_k-\bar{x}_{(i)}\right\}^2}{n-2}}$$

이므로 표본표준편차에 대한 표본영향함수 유도를 위해 식 (3.4)를 이용하여 다음과 같이 나타낸다.

$$s_{(i)}-s=\sqrt{\frac{-\frac{n}{n-1}(x_i-\bar{x})^2+(n-1)s^2}{n-2}}-s\tag{3.8}$$

그리고 미분을 통해 $s_{(i)}-s$가 갖는 함수적 특성을 살펴보기로 한다.

$$\frac{\partial\left\{s_{(i)}-s\right\}}{\partial x_i}=\frac{1}{2}\cdot\left\{\frac{-\frac{n}{n-1}(x_i-\bar{x})^2+(n-1)s^2}{n-2}\right\}^{-\frac{1}{2}}\cdot\frac{-2n\left(x_i-\bar{x}\right)}{(n-1)(n-2)}\tag{3.9}$$

이므로 $s_{(i)}-s$는 $\partial\{s_{(i)}-s\}/\partial x_i=0$이 성립하는 오직 $x_i=\bar{x}$에서만 극값을 갖는 초월함수(transcendental function)이다. 특히, 표본표준편차에 대한 경험적 영향함수는 식 (2.8)에 의해 $\mathrm{EIF}(s,x_i)=\{1/(2s)\}\cdot\{(x_i-\bar{x})^2-s^2\}$이므로, 이는 x_i에 대한 이차함수이자 함수 $s_{(i)}-s$와 동일한 위치인 $x_i=\bar{x}$에서 오직 하나의 극값을 갖는 함수임도 알 수 있다. 따라서 함수 $\mathrm{EIF}(s,x_i)$와 $\mathrm{SIF}(s,x_i)$가 동일하게 극값을 갖는 위치인 $x_i=\bar{x}$ 근방에서의 테일러 급수 전개(Taylor series expansion)에 의한 $\mathrm{SIF}(s,x_i)$의 2차 근사식과 이차함수인 경험적 영향함수 $\mathrm{EIF}(s,x_i)$와의 관계를 살펴 표본영향함수의 근사식을 유도하기로 한다. 이를 위하여 $s_{(i)}-s$은 다음과 같이 다시 나타낼 수 있다.

$$\sqrt{\frac{-\frac{n}{n-1}(x_i-\bar{x})^2+(n-1)s^2}{n-2}}-s\approx\sum_{k=1}^{3}a_k\left(x_i-\bar{x}\right)^{k-1}.\tag{3.10}$$

식 (3.10)의 양변에 $x_i = \bar{x}$를 대입하면 $a_1 = (\sqrt{(n-1)/(n-2)} - 1)s$임을 알 수 있다. 그리고 식 (3.10)의 양변을 x_i에 대해 미분하면

$$a_2 + 2a_3 (x_i - \bar{x}) = \frac{1}{2} \cdot \left\{ \frac{-\frac{n}{n-1} (x_i - \bar{x})^2 + (n-1)s^2}{n-2} \right\}^{-\frac{1}{2}} \cdot \frac{-2n (x_i - \bar{x})}{(n-1)(n-2)} \tag{3.11}$$

이므로 식 (3.11)의 양변에 $x_i - \bar{x}$를 대입하면 $a_2 = 0$을 얻는다. 마지막으로 식 (3.11)의 양변을 다시 x_i에 대해 미분하면

$$2a_3 = \frac{\frac{-n}{(n-1)(n-2)} \left\{ \frac{\frac{-n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2}{n-2} \right\} - \frac{n^2}{(n-1)^2(n-2)^2} (x_i - \bar{x})^2}{\left\{ \frac{\frac{-n}{n-1}(x_i - \bar{x})^2 + (n-1)s^2}{n-2} \right\}^{\frac{3}{2}}} \tag{3.12}$$

이고, 식 (3.12)의 양변에 $x_i = \bar{x}$를 대입하여 $a_3 = -n/\{2s(n-1)\sqrt{(n-1)(n-2)}\}$를 얻을 수 있다. 즉, $s_{(i)} - s$는 다음과 같이 테일러 2차 다항식으로 근사 가능하다.

$$s_{(i)} - s \approx -\frac{n}{2s(n-1)\sqrt{(n-1)(n-2)}} (x_i - \bar{x})^2 + \left(\sqrt{\frac{n-1}{n-2}} - 1 \right) s. \tag{3.13}$$

식 (3.13)는 $\mathrm{EIF}(s, x_i) = \{1/(2s)\} \cdot \{(x_i - \bar{x})^2 - s^2\}$를 이용해 다음과 같이 정리할 수 있다.

$$\begin{aligned} s_{(i)} - s &\approx -\frac{n}{2s(n-1)\sqrt{(n-1)(n-2)}} \left\{ (x_i - \bar{x})^2 - s^2 + s^2 \right\} + \left(\sqrt{\frac{n-1}{n-2}} - 1 \right) s \\ &= -\frac{n}{(n-1)\sqrt{(n-1)(n-2)}} \mathrm{EIF}(s, x_i) + s \left\{ \frac{(2n-1)\sqrt{n-2}}{2(n-1)\sqrt{n-1}} - 1 \right\}. \end{aligned} \tag{3.14}$$

따라서 표본표준편차에 대한 표본영향함수는 다음과 같이 표현할 수 있다.

$$\mathrm{SIF}(s, x_i) \approx \frac{n}{\sqrt{(n-1)(n-2)}} \mathrm{EIF}(s, x_i) + s \left\{ (n-1) - \frac{(2n-1)\sqrt{n-2}}{2\sqrt{n-1}} \right\}. \tag{3.15}$$

위에서 유도한 내용을 종합하여 표본평균, 표본분산, 표본표준편차에 대한 표본영향함수를 정리하면 다음과 같이 나타낼 수 있다. 특히, 표본평균에 대한 표본영향함수는 경험적 영향함수와 같지만, 표본분산과 표본표준편차에 대한 표본영향함수는 경험적 영향함수에 적당한 실수배와 함께 상수항의 합으로 표현이 됨을 알 수 있다.

$$\begin{aligned} \mathrm{SIF}(\bar{x}, x_i) &= \mathrm{EIF}(\bar{x}, x_i), \\ \mathrm{SIF}\left(s^2, x_i\right) &= \frac{n}{n-2} \mathrm{EIF}\left(s^2, x_i\right) + \frac{s^2}{n-2}, \\ \mathrm{SIF}(s, x_i) &\approx \frac{n}{\sqrt{(n-1)(n-2)}} \mathrm{EIF}(s, x_i) + s \left\{ (n-1) - \frac{(2n-1)\sqrt{n-2}}{2\sqrt{n-1}} \right\}. \end{aligned} \tag{3.16}$$

4. 모의실험을 통한 경험적 영향함수와 표본영향함수의 관계 확인

3장에서 표본평균, 표본분산, 표본표준편차에 대한 표본영향함수를 유도하고, 경험적 영향함수와 표본영향함수의 관계를 확인하였다. 4장에서는 3장에서 이론적으로 접근한 내용을 모의실험을 통해 경험적

Table 4.1. Summary of 300 random samples from $N(0,1)$

Min	Mean	Max	Variance
−2.30923	0.05943	2.78710	0.91174

Table 4.2. Summary of 300 samples shifted to be $\bar{x} = 4$

Min	Mean	Max	Variance
1.631	4.000	6.728	0.912

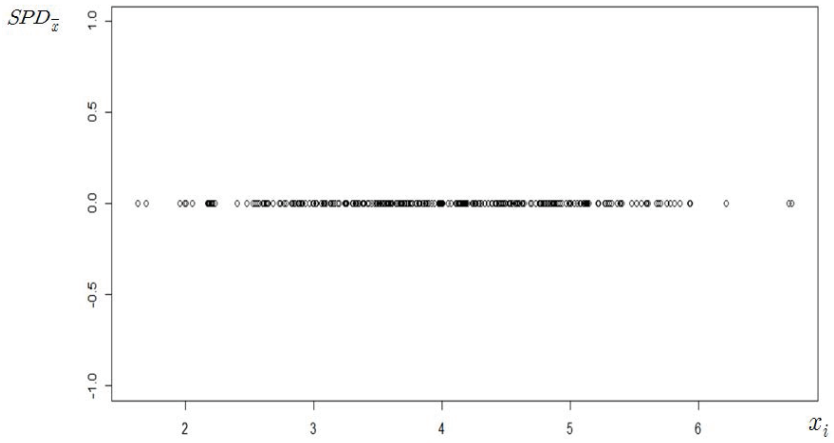

Figure 4.1. Graph of $\text{SPD}_{\bar{x}}$. SPD = simple prediction difference.

으로 확인하고자 한다. 모의실험을 진행하기 위해 R 통계 패키지에서 정규분포 $N(0,1)$을 따르는 임의추출한 300개의 표본을 사용하였고, 임의추출된 300개 표본의 기술 통계는 Table 4.1과 같다.

300개 표본의 표본평균이 0.05943이므로 일괄적으로 300개 데이터에서 0.05943을 빼고, 다시 4만큼을 더해서 $\bar{x} = 4$를 만족시키도록 보정한 300개 데이터에 대해 모의실험을 진행하였다. t통계량의 크기를 적절히 크게 하여 $t_{(i)} - t$의 값 변화의 관찰이 용이할 수 있게 하기 위해 $\bar{x} = 4$를 만족시키도록 시프트를 실시하였다. 초기에 생성한 데이터에서 $\bar{x} = 4$를 만족시키도록 시프트하여 다시 생성한 데이터의 기술 통계는 Table 4.2와 같다.

$\text{SIF}(T, x_i) \approx \text{EIF}(T, x_i)$에 의한 $T(\hat{F}_{(i)}) - T(\hat{F}) \approx \{-1/(n-1)\} \cdot \text{EIF}(T, x_i)$ 근사로 $T(\hat{F}_{(i)}) - T(\hat{F})$를 예측한 경우를 단순 근사(simple approximation; SA)라 하고, 단순 근사에 사용된 $\{-1/(n-1)\} \cdot \text{EIF}(T, x_i)$의 값을 SA_T로 표현하기로 한다. 또한, 단순 근사 과정에서 생긴 $T(\hat{F}_{(i)}) - T(\hat{F})$의 값과 $\{-1/(n-1)\} \cdot \text{EIF}(T, x_i)$의 값의 차이를 단순 예측 차이(simple prediction difference; SPD)라 하여 이 값을 SPD_T로 표현한다. 3장에서 유도한 $\text{EIF}(T, x_i)$에 실수배, 상수항의 합 보정을 한 식으로 $\text{SIF}(T, x_i)$에 대입 혹은 근사시켜 $T(\hat{F}_{(i)}) - T(\hat{F})$를 예측한 경우를 보정된 근사(calibrated approximation; CA)라 하고, CA_T로 표현한다. 보정된 근사에 의한 $T(\hat{F}_{(i)}) - T(\hat{F})$와 근사한 식의 차이를 보정한 예측 차이(calibrated prediction difference; CPD)라 하고 CPD_T로 나타내겠다.

4.1. $\bar{x}_{(i)} - \bar{x}$의 근사

단순 예측 차이는 $\text{SPD}_{\bar{x}} = (\bar{x}_{(i)} - \bar{x}) - [\{-1/(n-1)\} \cdot \text{EIF}(\bar{x}, x_i)]$이다. $\text{SPD}_{\bar{x}}$를 그래프로 나타내면 Figure 4.1과 같고, 모든 x_i에 대해 0의 값을 갖는다. 따라서 이를 통해 식 (3.1)의 결과인 $\text{SIF}(\bar{x}, x_i) =$

Table 4.3. Comparing differences in the approximation of $\bar{x}_{(i)} - \bar{x}$

ID	x_i	$\bar{x}$	$\bar{x}_{(i)} - \bar{x}$	$\mathrm{SA}_{\bar{x}}$	$\mathrm{SPD}_{\bar{x}}$
31	2.77011	4.00000	0.00411	0.00411	0.00000
32	2.78415	4.00000	0.00407	0.00407	0.00000
33	2.82730	4.00000	0.00392	0.00392	0.00000
			⋯		
101	3.57024	4.00000	0.00144	0.00144	0.00000
102	3.57436	4.00000	0.00142	0.00142	0.00000
103	3.58411	4.00000	0.00139	0.00139	0.00000
			⋯		
181	4.20198	4.00000	−0.00068	−0.00068	0.00000
182	4.23865	4.00000	−0.00080	−0.00080	0.00000
183	4.24454	4.00000	−0.00082	−0.00082	0.00000
			⋯		
261	5.08407	4.00000	−0.00363	−0.00363	0.00000
262	5.10051	4.00000	−0.00368	−0.00368	0.00000
263	5.10622	4.00000	−0.00370	−0.00370	0.00000

SA = simple approximation; SPD = simple prediction difference.

Table 4.4. Comparing differences in the approximation of $s^2_{(i)} - s^2$

ID	x_i	$s^2_{(i)} - s^2$	$\mathrm{EIF}(s^2, x_i)$	SA_{s^2}	SPD_{s^2}	CA_{s^2}	CPD_{s^2}
31	2.77011	−0.00204	0.60089	−0.00201	−0.00003	−0.00203	0.00000
32	2.78415	−0.00192	0.56655	−0.00189	−0.00003	−0.00192	0.00000
33	2.82730	−0.00158	0.46348	−0.00155	−0.00003	−0.00157	0.00000
				⋯			
101	3.57024	0.00243	−0.72704	0.00243	0.00000	0.00244	0.00000
102	3.57436	0.00245	−0.73056	0.00244	0.00001	0.00245	0.00000
103	3.58411	0.00247	−0.73877	0.00247	0.00000	0.00248	0.00000
				⋯			
181	4.20198	0.00292	−0.87094	0.00291	0.00001	0.00292	0.00000
182	4.23865	0.00286	−0.85478	0.00286	0.00000	0.00287	0.00000
183	4.24454	0.00285	−0.85194	0.00285	0.00000	0.00286	0.00000
				⋯			
261	5.08407	−0.00090	0.26348	−0.00088	−0.00002	−0.00090	0.00000
262	5.10051	−0.00102	0.29938	−0.00100	−0.00002	−0.00102	0.00000
263	5.10622	−0.00107	0.31199	−0.00104	−0.00003	−0.00106	0.00000

EIF = empirical influence function; SA = simple approximation; SPD = simple prediction difference; CA = calibrated approximation; CPD = calibrated prediction difference.

$\mathrm{EIF}(\bar{x}, x_i)$의 타당성을 확인할 수 있다. 또한 표본평균에 대한 표본영향함수를 근사시키기 위해서 경험적 영향함수를 대신 사용해도 동일한 결과를 얻을 수 있음을 알 수 있다. 모의실험에 사용한 데이터 x_i에 대한 $\bar{x}$, $\bar{x}_{(i)} - \bar{x}$, $\mathrm{SA}_{\bar{x}}$, $\mathrm{SPD}_{\bar{x}}$의 값은 각각 Table 4.3과 같다.

4.2. $s^2_{(i)} - s^2$의 근사

단순 예측 차이는 $\mathrm{SPD}_{s^2} = ({s^2}_{(i)} - s^2) - [\{-1/(n-1)\} \cdot \mathrm{EIF}(s^2, x_i)]$이고, 보정한 예측 차이는 $\mathrm{CPD}_{s^2} =$

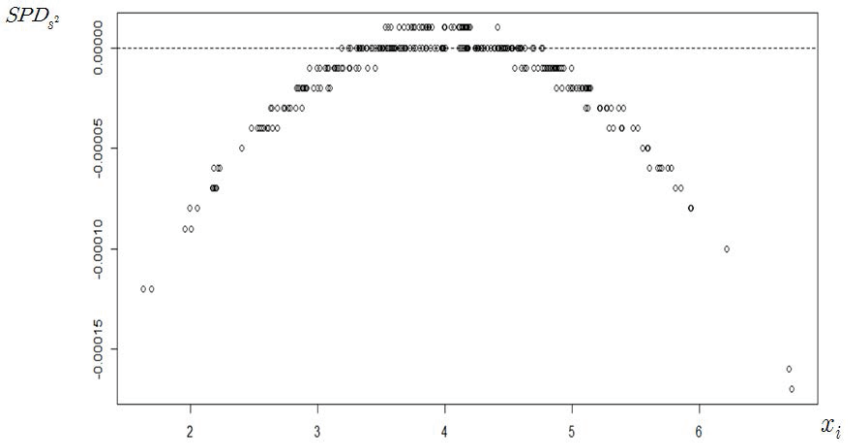

Figure 4.2. Graph of SPD_{s^2}. SPD = simple prediction difference.

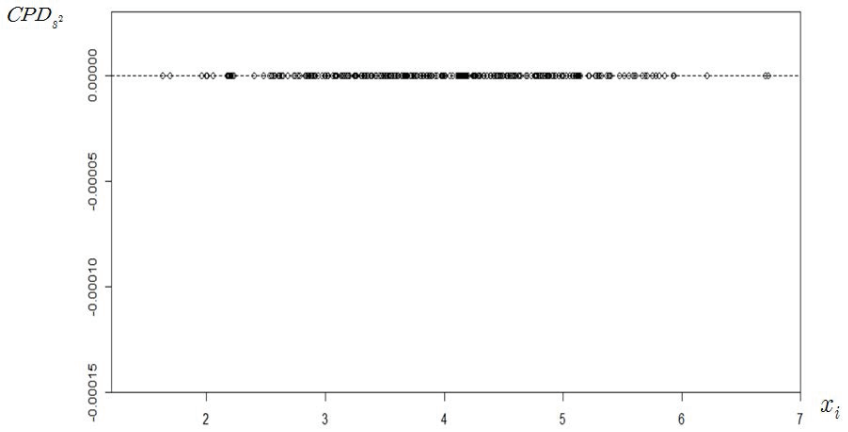

Figure 4.3. Graph of CPD_{s^2}. CPD = calibrated prediction difference.

$(s^2_{(i)} - s^2) - [-n/\{(n-1)(n-2)\} \cdot \mathrm{EIF}(s^2, x_i) - s^2/\{(n-1)(n-2)\}]$ 이다. Table 4.4에서 살펴보면 SPD_{s^2}의 경우 그 크기가 매우 작지만 오차가 발생하게 된다. SPD_{s^2}와 CPD_{s^2}의 그래프는 Figure 4.2, Figure 4.3과 같으며, SPD_{s^2}는 x_i의 양 끝 쪽으로 갈수록 그 크기가 증가하는 경향을 보인다. 한편, CPD_{s^2}의 값은 모든 x_i에 대해 0의 값을 갖고, 이는 $\mathrm{EIF}(s^2, x_i)$를 $\{n/(n-2)\}\cdot\mathrm{EIF}(s^2, x_i) + s^2/(n-2)$로 보정하여 $\mathrm{SIF}(s^2, x_i)$를 근사하는 것이 타당함을 설명한다.

4.3. $s_{(i)} - s$의 근사

단순 예측 차이는 $\mathrm{SPD}_s = \{s_{(i)} - s\} - [\{-\{1/(n-1)\} \cdot \mathrm{EIF}(s, x_i)\}]$이고, 보정한 예측 차이는 $\mathrm{CPD}_s = \{s_{(i)} - s\} - [-n\mathrm{EIF}(s, x_i)/\{(n-1)\sqrt{(n-1)(n-2)}\} + s\{\{(2n-1)\sqrt{n-2}\}/\{2(n-1)\sqrt{n-1}\} - 1\}]$이다. Figure 4.4, Figure 4.5와 Table 4.5를 살펴보면 SPD_s와 CPD_s 모두 x_i의 양 끝 쪽으로 갈수록 그 크기가 증가하지만 CPD_s는 SPD_s에 비해 0의 값에 수렴하는 비율이 높고, 그 평균은 0이다. 즉, $\mathrm{SIF}(s, x_i)$는 $\{n/\sqrt{(n-1)(n-2)}\} \cdot \mathrm{EIF}(s, x_i) + s[(n-1) - \{(2n-1)\sqrt{n-2}/(2\sqrt{n-1})\}]$로 보정해

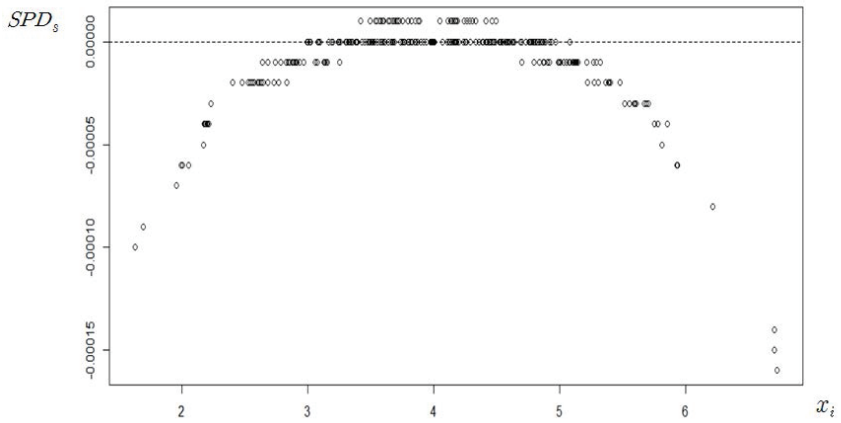

Figure 4.4. Graph of SPD_s. SPD = simple prediction difference.

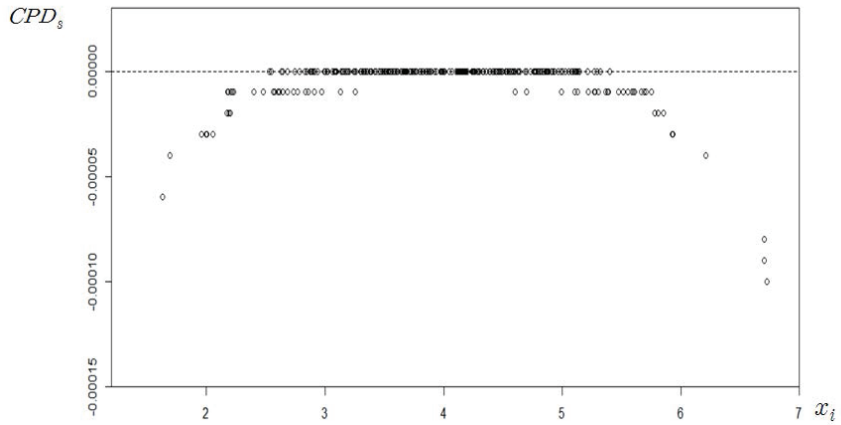

Figure 4.5. Graph of CPD_s. CPD = calibrated prediction difference.

근사하여 예측하는 것이 보다 정확도가 높다고 할 수 있다.

5. 실제 자료 분석 과정에의 적용

사회과학 연구에서 폭넓게 활용되는 t통계량은 표본평균 $\bar{x}$, 검정하고 싶은 값 μ_0, 표본표준편차 s, 그리고 표본의 크기 n에 대하여 $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$로 얻어진다. 이때, 추출된 표본의 관측값으로부터 얻어진 $\bar{x}$와 s는 이상치가 포함되는 경우 그 값의 심각한 변동이 있을 수 있고, 이는 t통계량에도 위협이 될 수 있다. 본 연구에서 유도하고 타당성을 검증한 표본영향함수의 근사 방법을 이용하여 실제 자료 분석 과정에서 표본평균, 표본분산, 표본표준편차에서 이상치의 제거를 고려해야하는 경우, 제거 순위를 결정하는 방법을 적용한 예를 살펴본다. 실제 자료 분석에는 2020년 대전 지역의 한 고등학교 3학년 학생 226명의 수학 점수 자료를 활용하였으며, 226명의 수학 점수에 대한 기술 통계는 Table 5.1과 같다. 사회과학 연구 수행을 위해 20명의 점수를 임의추출한 연구 상황을 가정하였다. R을 이용해 20명의 점수를 임의로 추출한 뒤, 각각의 관측값이 표본평균, 표본분산, 표본표준편차에 미치는 영향을 본 연구에서 제안한 보정된 표본영향함수 근사의 방법으로 얻은 값의 절댓값으로 산출하였다. 그리고 이

Table 4.5. Comparing differences in the approximation of $s_{(i)} - s$

ID	x_i	$s_{(i)} - s$	EIF(s, x_i)	SA$_s$	SPD$_s$	CA$_s$	CPD$_s$
31	2.77011	−0.00107	0.31465	−0.00105	−0.00002	−0.00106	−0.00001
32	2.78415	−0.00100	0.29667	−0.00099	−0.00001	−0.00100	0.00000
33	2.82730	−0.00082	0.24270	−0.00081	−0.00001	−0.00082	0.00000
				⋯			
101	3.57024	0.00128	−0.38071	0.00127	0.00001	0.00128	0.00000
102	3.57436	0.00128	−0.38255	0.00128	0.00000	0.00128	0.00000
103	3.58411	0.00130	−0.38685	0.00129	0.00001	0.00130	0.00000
				⋯			
181	4.20198	0.00153	−0.45606	0.00153	0.00000	0.00153	0.00000
182	4.23865	0.00150	−0.44760	0.00150	0.00000	0.00150	0.00000
183	4.24454	0.00150	−0.44611	0.00149	0.00001	0.00150	0.00000
				⋯			
261	5.08407	−0.00047	0.13797	−0.00046	−0.00001	−0.00047	0.00000
262	5.10051	−0.00053	0.15677	−0.00052	−0.00001	−0.00053	0.00000
263	5.10622	−0.00056	0.16337	−0.00055	−0.00001	−0.00055	−0.00001

EIF = empirical influence function; SA = simple approximation; SPD = simple prediction difference; CA = calibrated approximation; CPD = calibrated prediction difference.

Table 5.1. Influence of observation and outlier removal ranking

Mean	Median	Variance	Standard deviation
55.23	56.70	322.0569	17.9459

값이 클수록 각 통계량에 미치는 영향이 큰 것으로 판단할 수 있으므로 통계량의 변화에 미치는 영향이 큰 이상치 제거를 고려해야할 경우, 이 값이 상대적으로 높은 순서대로 제거 순위를 부여하는 방법으로써 이상치 제거 기준을 세울 수 있다. 추출된 20명의 수학 점수의 표본평균과 표본표준편차는 각각 $\bar{x} = 60.6$, $s = 18.6946$이었으며, 표본평균, 표본분산, 표본표준편차에 영향이 큰 이상치의 제거 순위를 Table 5.2와 같이 제안하였다.

6. 결론

본 연구에서는 섭동을 고려한 식 $T(\hat{F}_{(i)}) - T(\hat{F}) = \{-1/(n-1)\} \cdot \mathrm{SIF}(T, x_i)$에서 표본영향함수 $\mathrm{SIF}(T, x_i)$를 경험적 영향함수 $\mathrm{EIF}(T, x_i)$로 근사시키는 방법에서 정확성을 높이기 위해 논의한 본 연구의 결론은 다음과 같다.

첫째, $\mathrm{SIF}(\bar{x}, x_i)$와 $\mathrm{EIF}(\bar{x}, x_i)$는 동일하므로 $\mathrm{SIF}(\bar{x}, x_i)$의 근사를 위해 $\mathrm{EIF}(\bar{x}, x_i)$를 대신 사용해도 되며, 이때 오차는 발생하지 않는다. $\bar{x}_{(i)} - \bar{x} = \{-1/(n-1)\}\mathrm{SIF}(T, x_i)$의 타당성을 확인하였다.

둘째, $\mathrm{SIF}(s^2, x_i)$에 대해 정확도 높은 근사를 위해서는 $\mathrm{EIF}(s^2, x_i)$에 적당한 실수배와 상수항의 합으로 보정이 필요하다. 수리적으로 $\mathrm{SIF}(s^2, x_i) = \{n/(n-2)\} \cdot \mathrm{EIF}(s^2, x_i) + s^2/(n-2)$의 식이 성립함을 보였고, $s^2_{(i)} - s^2 = \{-n/(n-1)(n-2)\} \cdot \mathrm{EIF}(s^2, x_i) - \{s^2/(n-1)(n-2)\}$와 같이 근사하면 예측 차이가 0이 되는 것을 모의실험으로 관찰하여 그 타당성을 확인하였다.

셋째, $\mathrm{SIF}(s, x_i)$를 $\mathrm{EIF}(s, x_i)$로 근사시키기 위해서는 실수배와 상수항의 합 보정이 필요하다. 수리적으로 $\mathrm{SIF}(s, x_i) \approx n/\sqrt{(n-1)(n-2)} \cdot \mathrm{EIF}(s, x_i) + s[(n-1) - \{(2n-1)\sqrt{n-2}\}/(2\sqrt{n-1})]$이 성립함을 보였고, $s_{(i)} - s$의 예측에서 단순 근사시키는 방법에 비해 상대적으로 정확성이 높음을 알 수 있었

Table 5.2. Influence of observation and outlier removal ranking

ID	x_i (score)	$\bar{x}_{(i)}$	$\lvert SA_{\bar{x}}\rvert$	Removal rank ($\bar{x}$)	$s_{(i)}$	$\lvert CA_{s^2}\rvert$	Removal rank (s^2)	$\lvert CA_s\rvert$	Removal rank (s)
1	94.6	58.8105	1.7895	3	17.3581	48.1863	3	1.2476	3
9	81.1	59.5211	1.0789	4	18.5561	5.1600	18	0.1275	18
23	75.1	59.8368	0.7632	7	18.8841	7.1207	17	0.1922	17
28	75.0	59.8421	0.7579	8	18.8886	7.2897	16	0.1966	16
29	74.9	59.8474	0.7526	9	18.8930	7.4575	15	0.2010	15
38	70.5	60.0789	0.5211	10	19.0571	13.6844	14	0.3631	14
42	70.4	60.0842	0.5158	11	19.0601	13.7996	13	0.3661	13
46	70.2	60.0947	0.5053	12	19.0661	14.0265	12	0.3720	12
53	70.0	60.1053	0.4947	13	19.0719	14.2488	11	0.3778	11
66	66.2	60.3053	0.2947	16	19.1591	17.5821	8	0.4645	8
87	65.2	60.3579	0.2421	18	19.1746	18.1786	6	0.4801	6
122	56.4	60.8211	0.2211	19	19.1800	18.3844	4	0.4854	4
123	56.4	60.8211	0.2211	19	19.1800	18.3844	4	0.4854	4
132	55.8	60.8526	0.2526	17	19.1718	18.0686	7	0.4772	7
141	51.6	61.0737	0.4737	15	19.0832	14.6792	9	0.3890	9
143	51.5	61.0789	0.4789	14	19.0804	14.5733	10	0.3862	10
178	42.3	61.5632	0.9632	6	18.6901	0.1682	20	0.0025	20
180	42.1	61.5737	0.9737	5	18.6786	0.5986	19	0.0088	19
209	25.5	62.4474	1.8474	2	17.2295	52.6314	2	1.3633	2
220	17.2	62.8842	2.2842	1	16.0859	90.7337	1	2.3552	1

SA = simple approximation; CA = calibrated approximation.

다. $s_{(i)}-s \approx [-n/\{(n-1)\sqrt{(n-1)(n-2)}\}]\cdot \mathrm{EIF}\,(s,x_i)+s[\{(2n-1)\sqrt{n-2}\}/\{2(n-1)\sqrt{n-1}\}-1]$의 타당성을 확인하였다.

넷째, 본 논문에서 제안한 경험적 영향함수의 보정을 이용한 표본영향함수의 근사 방법이 실제 자료 분석 과정에서도 적용될 수 있으며, 이를 통한 이상치 선정이 가능함을 볼 수 있었다.

References

Campbell, N. A. (1978). The influence function as an aid outlier detection in discrimination 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 and empirical influence function for detection influential cases in regression, *Technometrics*, **22**, 495–508.

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

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

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

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

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

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

Kim, S. and Kim, H. (2019). A study on the performance of the influence function on the t-statistic

depending on population distributions, *Journal of the Korean Data & Information Science Society*, **30**, 573–585.

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

Lee, H. and Kim, H. (2008). Influence function on the coefficient of variation, *Communications for Statistical Applications and Methods*, **15**, 509–516.

Park, S. and Kim, H. (2019). A study on the location of the observation which has the least effect on the t-statistic, *Journal of the Korean Data & Information Science Society*, **30**, 1221–1232.

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

경험적 영향함수와 표본영향함수의 차이 및 보정에 관한 연구

강현석[a] · 김홍기[b,1]

[a]대전고등학교, [b]충남대학교 정보통계학과

(2020년 5월 20일 접수, 2020년 6월 18일 수정, 2020년 7월 3일 채택)

요 약

이상치에 대한 적절한 선별과 배제없이 모든 데이터를 종합적으로 분석하게 되는 경우 데이터 분석을 통해 얻은 결과의 신뢰성과 해석의 일반성에 치명적인 위협을 받을 수 있다. 따라서 데이터의 분석 과정에서 이러한 이상치를 판별하고, 이상치가 통계량, 통계적 모형에 어떠한 영향을 주는 지에 대한 분석은 매우 중요한 일이라 할 수 있다. Hampel이 영향함수를 활용하여 이상치를 판별할 수 있는 방법을 소개한 이후, 이상치를 판별하기 위한 방법론으로 영향함수가 폭넓게 활용되어 왔다. 영향함수에는 경험적 영향함수와 표본영향함수가 있으며, 경험적 영향함수를 활용해 표본영향함수를 근사 추론하여 하나의 관측값이 제거되었을 때 통계량에 미치는 영향을 예측하는 방법론이 주로 활용되었다. 본 연구에서는 표본평균, 표본분산, 표본표준편차의 표본영향함수 유도를 통해 경험적 영향함수와 표본영향함수의 차이를 살펴 본다. 또한 경험적 영향함수로 표본영향함수를 근사하는 과정에서 발생하는 오차를 줄이기 위해 경험적 영향함수의 보정으로 표본영향함수를 근사 추론하는 방법을 제안하고, 모의실험을 통해 제안한 추론 방법의 타당성을 확인한다.

주요용어: 영향함수, 이상치, 경험적 영향함수, 표본영향함수

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