The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.6.953
26(6), 953–958

Measurement Error Model with Skewed Normal Distribution

Tae-Young Heo[a] · Jungsoon Choi[b] · Man Sik Park[c,1]

[a]Department of Information Statistics, Chungbuk National University
[b]Department of Mathematics, Hanyang University
[c]Department of Statistics, Sungshin Women's University

(Received August 30, 2013; Revised November 6, 2013; Accepted November 18, 2013)

Abstract

This study suggests a measurement error model based on skewed normal distribution instead of normal distribution to identify slope parameter properties in a simple liner regression model. We prove that the slope parameter in a simple linear regression model is underestimated.

Keywords: Measurement error model, skewed normal distribution, consistency, simulation.

1. 서론

일반적인 회귀분석에서와는 달리 설명변수의 값을 관찰하는데 오차가 수반된다는 가정 하에서의 모형을 측정오차모형이라고 하며 이 경우 일반적인 선형회귀분석에서의 최소제곱추정량은 편의추정량이고 일치성을 유지하지 못한다. 측정오차 모형에서 설명변수에 수반되는 오차는 일반적으로 정규분포를 가정한다. 본 연구에서는 관측되는 독립변수의 변동성이 정규분포가 아닌 왜도정규분포(skewed normal distribution; SN)에 기반을 둔 측정오차모형을 제시하고 단순선형회귀모형에서의 기울기 계수에 대하여 그 특성을 파악하였다.

일반적인 일변량(univariate) 선형측정오차모형은 관측되지 않은 자료 $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$와 $\mathbf{w} = (W_1, W_2, \ldots, W_n)^T$에 다음과 같은 모형을 가정한다 (Kendall과 Stuart, 1979; Fuller, 1987; Cheng과 Van Ness, 1999).

$$\boldsymbol{\mu} = \alpha\mathbf{1} + \beta\mathbf{w}.$$

여기서, $\mathbf{1}$는 n크기의 열벡터(column vector)이고 α와 β는 추정해야 할 모수이다. 현실적으로 관측가능한 자료, $\mathbf{x} = (X_1, \ldots, X_n)^T$와 $\mathbf{y} = (Y_1, \ldots, Y_n)^T$은 각각 다음과 같이 표현되어질 수 있다. 임의의 $i = 1, 2, \ldots, n$에 대하여

$$X_i = W_i + \epsilon_i,$$
$$Y_i = \mu_i + \nu_i.$$

This research was supported by the research fund of Chungbuk National University in 2012.
[1]Corresponding author: Assistant Professor, Department of Statistics, Sungshin Women's University, 249-1 Dongseon-Dong 3-Ga, Seongbuk-Gu, Seoul 136-742, Korea. E-mail: mansikpark@sungshin.ac.kr

여기서, 오차항인 ϵ_i와 ν_i는 평균이 모두 0이고 유한한 분산을 가지고 있으며, 일반적으로 이변량 정규분포를 따르는 벡터로 가정한다.

왜도정규분포(SN)에서는 분포함수(distribution function)에 왜도(skewness)의 정도를 의미하는 형상모수(shape parameter)를 포함시켰으며, 이는 정규분포의 확장된 형태라고 할 수 있다. 만약 확률변수, ϵ이 왜도정규분포를 따른다면 확률밀도함수(probability density function)는 다음과 같다.

$$f(\epsilon; a, b, \delta) = \frac{2}{b}\phi\left(\frac{\epsilon - a}{b}\right) \Phi\left(\delta\frac{\epsilon - a}{b}\right).$$

여기서, ϕ와 Φ는 표준정규분포(standard normal distribution)의 확률밀도함수와 분포함수이며, a와 b는 각각 위치모수(location parameter)와 척도모수(scale parameter), 그리고 δ는 형상모수를 의미한다. 만약 δ가 0으로 접근하면 정규분포와 유사하며, δ의 절대값이 클수록 왜도의 정도가 커지게 된다. 따라서, 주어진 확률변수, ϵ이 왜도정규분포를 따른다고 하면 $\epsilon \sim SN\left(a, b^2; \delta\right)$라고 표현한다. 확률변수, ϵ의 평균과 분산은 다음과 같다 (Azzalini, 1985).

$$\mathrm{E}(\epsilon) = a + b\sqrt{\frac{2\delta^2}{\pi\left(1+\delta^2\right)}}, \quad \mathrm{Var}(\epsilon) = b^2\left(1 - \frac{2\delta^2}{\pi\left(1+\delta^2\right)}\right). \tag{1.1}$$

이 논문에서는 측정오차 모형에서 오차항의 분포를 왜도정규분포로 가정하는 경우에 대해 연구하고자 한다. 본 논문의 순서는 다음과 같다. 제 2장에서는 왜도정규분포를 고려한 측정오차 모형에 대해 기술하고자 한다. 제 3장에서는 오차항이 왜도정규분포를 따른다는 가정 하에서의 단순선형회귀모형에 대해 기술하고 제 4장에서는 왜도의 정도에 따라 단순선형회귀모형 하에서의 기울기 계수가 어떤 특성을 가지게 되는지를 파악하고자 한다. 마지막으로, 결론은 제 5장에 기술하고자 한다.

2. 왜도정규분포를 고려한 측정오차 모형

2.1. 모형 정의 및 적률

관측가능한 독립변수, X_i는 오차가 존재하지 않는 상수, W_i와 왜도정규분포를 따르는 오차항의 선형결합으로 표현되고 이러한 가정 하에서의 측정오차모형은 다음과 같이 정의할 수 있다. 임의의 $i = 1, 2, \ldots, n$에 대하여

$$X_i = W_i + \sigma_W c_i. \tag{2.1}$$

여기서, σ_W는 임의의 상수이다. 오차항인 c_i는 다음과 같이 가정하기로 한다.

$$c_i = \lambda p_i + \gamma|q_i|.$$

여기서, p_i와 q_i는 표준정규분포를 따르는, 서로 독립인 확률변수이며 c_i의 확률밀도함수는 다음과 같이 구할 수 있다.

$$\begin{aligned} f(c_i; \lambda, \gamma) &= \int_{-\infty}^{\infty} \frac{1}{\lambda}\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(c_i - \gamma|q_i|)^2}{2\lambda^2}\right\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{q_i^2}{2}\right\} dq_i \\ &= \frac{\lambda^{-1}}{2\pi} \exp\left\{-\frac{c_i^2}{2(\lambda^2+\gamma^2)}\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{\lambda^2+\gamma^2}{2\pi\lambda^2}\left(|q_i| - \frac{\gamma c_i}{\lambda^2+\gamma^2}\right)^2\right\} dq_i \\ &= \frac{2}{\sqrt{\lambda^2+\gamma^2}}\phi\left(\frac{c_i}{\sqrt{\lambda^2+\gamma^2}}\right) \Phi\left(\frac{\gamma}{\lambda}\frac{c_i}{\sqrt{\lambda^2+\gamma^2}}\right). \end{aligned} \tag{2.2}$$

따라서, 확률변수 c_i는 다음과 같은 왜도정규분포를 따르게 된다. 임의의 $i = 1, 2, \ldots, n$에 대하여

$$c_i \sim SN\left(0, \lambda^2 + \gamma^2; \frac{\gamma}{\lambda}\right).$$

여기서, 임의의 실수인 λ, γ에 대하여, $\sqrt{\lambda^2 + \gamma^2}$는 척도모수이고 γ/λ는 형상모수로서 왜도의 정도를 나타내는 모수이다. 오차가 존재하지 않는 상수에 대해 $W_i \sim N(\omega, \sigma_W^2)$라고 가정한다면 관측가능한 독립변수인 X_i는 식 (2.2)로부터 다음과 같은 확률밀도함수를 갖게 된다.

$$f(x_i; \lambda, \gamma, \omega, \sigma_W) = \frac{2}{\sigma_W\sqrt{\lambda^2 + \gamma^2 + 1}}\phi\left(\frac{x_i - \omega}{\sigma_W\sqrt{\lambda^2 + \gamma^2 + 1}}\right)\Phi\left(\frac{\gamma}{\sqrt{\lambda^2 + 1}}\frac{x_i - \omega}{\sigma_W\sqrt{\lambda^2 + \gamma^2 + 1}}\right)$$

따라서, X_i는 다음과 같은 왜도정규분포의 형태를 가지게 된다.

$$X_i \sim SN\left(\omega, \sigma_W^2(\lambda^2 + \gamma^2 + 1); \frac{\gamma}{\sqrt{\lambda^2 + 1}}\right). \tag{2.3}$$

식 (1.1)에 의해 X_i의 평균과 분산은 다음과 같이 계산된다.

$$\mathrm{E}(X_i) = \omega + \sigma_W\sqrt{\frac{2\gamma^2}{\pi}}, \quad \mathrm{Var}(X_i) = \sigma_W^2\left\{\lambda^2 + \left(1 - \frac{2}{\pi}\right)\gamma^2 + 1\right\}.$$

2.2. 왜도정규분포 하에서의 단순선형회귀모형

다음과 같은 단순선형회귀모형이 있다고 가정하자.

$$Y_i = \alpha + \beta W_i + \nu_i. \tag{2.4}$$

여기서, 모든 $i = 1, \ldots, n$에 대해 $\nu_i \sim N(0, \sigma_\nu^2)$이다. 식 (2.3)과 같은 왜도정규분포를 고려하지 않은 상태에서의 회귀계수 β에 대한 최소제곱추정량은 다음과 같이 정의된다.

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(W_i - \bar{W}\right)}{\sum_{i=1}^{n}\left(W_i - \bar{W}\right)^2}$$

여기서, $\bar{Y}$와 $\bar{W}$는 두 확률변수 각각의 표본평균을 의미하며, 최소제곱추정량 $\widehat{\beta}$은 회귀계수 β의 최소분산불편추정량이다. 두 확률변수 W와 Y의 제곱합(S_{WY})과 W의 제곱합(S_W^2)을 구하면 다음과 같다.

$$S_{WY} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(W_i - \bar{W}\right), \quad S_W^2 = \sum_{i=1}^{n}\left(W_i - \bar{W}\right)^2.$$

그리고 식 (2.4) 하에서 최소제곱추정량 $\widehat{\beta}$은 다음과 같은 일치성(consistency)을 만족하게 된다.

$$\widehat{\beta} = \frac{S_{WY}}{S_W^2} \xrightarrow{p} \beta = \frac{\sigma_{WY}}{\sigma_W^2}.$$

반면에 식 (2.4)에서 표현된 정규분포의 형태를 취하는 W_i 대신 왜도정규분포의 형태를 취하는 X_i를 가정하면, 식 (2.1)을 통해 다음과 같은 모형을 가정할 수 있다.

$$Y_i = \alpha + \beta X_i + \nu_i - \beta\sigma_W c_i.$$

Table 3.1. Estimation Results of the Slope Parameter, β when $X_i \sim N(0,1)$

a	b	σ_W^2	[distribution]	β			
				Mean	Bias	Variance	MSE
1	2	a^2+b^2	[Normal]	0.341	−1.659	0.0023	2.754
1	2	a^2+b^2	$[SN]$	0.592	−1.408	0.00416	1.986
1	5	a^2+b^2	[Normal]	0.074	−1.926	0.000845	3.710
1	5	a^2+b^2	$[SN]$	0.180	−1.820	0.00197	3.314
1	2	$a^2+(1-2/\pi)b^2$	[Normal]	0.583	−1.417	0.00392	2.012
1	2	$a^2+(1-2/\pi)b^2$	$[SN]$	0.584	−1.416	0.00424	2.010
1	5	$a^2+(1-2/\pi)b^2$	[Normal]	0.183	−1.817	0.0019	3.302
1	5	$a^2+(1-2/\pi)b^2$	$[SN]$	0.184	−1.816	0.00197	3.298

여기서, 모든 $i=1,\ldots,n$에 대하여 ν_i와 c_i는 서로 독립이라고 가정한다. 왜도정규분포를 고려한 측정오차모형에서의 회귀계수 추정량, $\widetilde{\beta}$은 다음과 같이 표현된다.

$$\widetilde{\beta} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}.$$

여기서, $\bar{X}$는 X의 표본평균이다. 두 확률변수 X와 Y의 제곱합(S_{XY})을 구하면 다음과 같다.

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right) = \sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)X_i \\ &= \sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(W_i+\sigma_W c_i\right) = S_{WY}+\sigma_W S_{CY}. \end{aligned}$$

여기서, $S_{CY}=\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)(c_i-\bar{c})$는 두 확률변수 c와 Y의 제곱합이며, c의 표본평균이 $\bar{c}$이다. 왜도정규분포 하에서의 $\widetilde{\beta}$는 다음과 같은 확률근사(convergence in probability)를 보이게 된다.

$$\widetilde{\beta} = \frac{S_{XY}}{S_X^2} \xrightarrow{p} \frac{\sigma_{WY}}{\sigma_X^2} = \beta\left\{\lambda^2+\left(1-\frac{2}{\pi}\right)\gamma^2+1\right\}^{-1} \le \beta.$$

따라서 왜도정규분포를 따르는 측정오차항을 가진 X_i와 c_i와 ν_i가 독립인 경우, 회귀계수의 최소제곱 추정량, $\widetilde{\beta}$는 편의되어 있다는 것을 알 수 있다. 즉, 회귀계수는 측정오차로 인해 $\{\lambda^2+(1-2/\pi)\gamma^2+1\}^{-1}$만큼 과소추정되며, 과소추정의 정도는 W_i의 분산인 σ_W^2에 의존하지 않음을 알 수 있다.

3. 모의실험

본 모의실험에서는 측정오차 모형 하에서의 측정오차에 의한 회귀계수의 과소추정 정도를 파악하기 위하여 정규분포에서의 $\widehat{\beta}$와 왜도정규분포에서의 $\widetilde{\beta}$을 비교하였다. 모의실험에 사용된 계수는 $\beta=2$이며 X_i는 $X_i\sim N(0,1)$과 $X\sim N(0,4)$이다. 왜도정규분포를 위해 $a=1$, $b=2$, 5, 그리고 $\sigma_W^2=a^2+b^2$, $a^2+(1-2/\pi)b^2$를 이용하였다. 모의실험 조건에서 $a=1$, $b=2$, $a^2+b^2=5$이면 정규분포 기반의 측정오차 모형에서 $\widetilde{\beta}$은 $2/(1+5)=0.333$으로 확률적으로 수렴하며, 왜도정규분포 기반의 측정오차에서는 $2/(1+1+4(1-2/\pi))=0.579$로 수렴한다.

$X_i\sim N(0,1)$인 경우 Table 3.1을 통해, $\sigma_W^2=a^2+b^2$인 경우 Bias 측면에서 정규분포를 기반으로 한 측정오차 모형보다는 왜도정규분포 기반의 측정오차 모형의 추정값이 더 좋음을 알 수 있다. 반면,

Table 3.2. Estimation Results of the Slope Parameter, β when $X_i \sim N(0,4)$

a	b	σ_W^2	[distribution]	β			
				Mean	Bias	Variance	MSE
1	2	a^2+b^2	[Normal]	0.325	−1.675	0.00223	2.806
1	2	a^2+b^2	$[SN]$	0.567	−1.432	0.00311	2.056
1	5	a^2+b^2	[Normal]	0.083	−1.917	0.000789	3.675
1	5	a^2+b^2	$[SN]$	0.203	−1.797	0.00171	3.231
1	2	$a^2+(1-2/\pi)b^2$	[Normal]	0.620	−1.380	0.00289	1.908
1	2	$a^2+(1-2/\pi)b^2$	$[SN]$	0.621	−1.380	0.00313	1.906
1	5	$a^2+(1-2/\pi)b^2$	[Normal]	0.178	−1.822	0.00153	3.321
1	5	$a^2+(1-2/\pi)b^2$	$[SN]$	0.179	−1.821	0.00158	3.319

$\sigma_W^2 = a^2+(1-2/\pi)b^2$인 경우는 정규분포와 왜도정규분포간 Bias 차이가 크지 않음을 알 수 있다. 그리고, 왜도정규분포 기반의 측정오차 모형의 MSE 값이 정규분포 기반의 측정오차 모형보다 훨씬 작은 값을 가진다. 따라서, MSE 측면에서는 왜도정규분포 기반의 측정오차 모형이 월등히 좋음을 알 수 있다.

Table 3.2에서는 $X_i \sim N(0,4)$인 경우 정규분포와 왜도정규분포 기반의 측정오차 모형의 기울기 계수를 비교하였다. 마찬가지로, $\sigma_W^2 = a^2+b^2$일 때 왜도정규분포 기반의 측정오차 모형이 Bias 측면에서 더 좋은 성능을 보임을 알 수 있으며, $\sigma_W^2 = a^2+(1-2/\pi)b^2$일 때는 두 분포간 차이가 없음을 알 수 있다. 또한, 왜도정규분포 기반의 측정오차 모형에서 추정된 기울기 계수가 더 작은 MSE값을 가지므로, MSE 측면에서 왜도정규분포 기반의 측정오차 모형이 참값을 더 잘 추정함을 알 수 있다.

4. 결론

본 연구에서는 정규분포 기반이 아닌 왜도정규분포 기반의 측정오차모형을 제시하고 단순선형회귀모형에서 기울기 계수에 대해 특성을 파악하였다. 측정오차를 가진 단순선형회귀모형의 회귀계수는 측정오차로 인해 최소제곱추정량은 편의되어 과소추정 되고 있음을 알 수 있었다. 다양한 조건에서의 모의실험을 통해 측정오차에 의해 회귀계수의 감쇠 정도를 파악하였다.

References

Azzalini, A. (1985). A Class of Distribution Which Includes the Normal Ones, *Scandinavian Journal of Statistics*, **12**, 171–178.

Cheng, C. L. and Van Ness, J. W. (1999). *Statistical Regression with Measurement Error*, Arnold, New York.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, New York.

Kendall, M. G. and Stuart, A. (1979). *The Advanced Theory of Statistics*, Griffin, London.

왜도정규분포 기반의 측정오차모형

허태영[a] · 최정순[b] · 박만식[c,1]

[a]충북대학교 정보통계학과, [b]한양대학교 수학과, [c]성신여자대학교 통계학과

(2013년 8월 30일 접수, 2013년 11월 6일 수정, 2013년 11월 18일 채택)

요 약

본 연구에서는 정규분포 기반이 아닌 왜도정규분포 기반의 측정오차모형을 제시하고 단순선형회귀모형에서의 기울기 계수에 대하여 특성을 파악하였다. 모의실험을 통해 측정오차모형에서 단순선형회귀모형에서의 기울기 계수의 과소추정 및 감쇠의 정도를 보였다.

주요용어: 측정오차모형, 왜도정규분포, 일치성, 모의실험.

이 논문은 2012년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음.

[1]교신저자: (136-742) 서울 성북구 동선동 3가 249-1, 성신여자대학교 자연과학대학 통계학과, 조교수.
E-mail: mansikpark@sungshin.ac.kr