한국통계학회논문집 제2권 1호
1995년 4월 pp. 13-21

선형모형에서 오차의 대칭성에 대한 검정과 회귀계수의 추정에 관한 연구[1)]

김 순 옥[2)]

요 약

선형모형에서 오차가 대칭인 분포를 따르는지 또는 한쪽으로 치우친(skewed distribution)분포를 따르는지 검정하는 문제를 다루었다. 또 이러한 검정과정을 분석의 예비단계로 하는 회귀계수의 추정방법에 대해서 연구하고, 모의실험을 통해서 회귀계수 추정법들의 효율을 비교하였다.

1. 서론

선형모형

$$Y = \beta_o + \beta_1 x + \varepsilon$$

에서 오차분포의 대칭성여부는 최소제곱법뿐만 아니라 로버스트한 방법을 적용할 때도 관심의 대상이 된다. 한쪽으로 치우친 분포(skewed distribution)는 물리학이나 공학분야 등에서 아주 높거나 아주 낮은 온도, 홍수수위, 인장강도와 같이 극단치를 다루는 과정에서 흔히 발생하고, 이들을 반응변수로 회귀분석 기법을 적용할 때는 분포의 대칭성여부에 주의를 기울일 필요가 있다. 반응변수의 분포, 즉 오차의 분포,가 한쪽으로 치우쳐있을 경우에 최소제곱법은 물론이고 모집단 분포의 대칭성을 가정하는 로버스트통계량들의 효율성도 떨어진다. 따라서 오차분포의 대칭성 여부를 탐지해 내는 것은 모형을 설정하고 추론하는데 있어서 중요한 과정이라 생각된다.

본 연구에서는 오차분포가 대칭인지, 한쪽으로 치우쳤는지를 검정하고 이에 따른 추정법을 제안하려고 한다. 오차에 대한 검정은 잔차($e = Y - \hat{Y}$)를 이용한다. 2절에서 일변량 분포의 대칭성에 관한 검정법들을 다루었고, 3절에서는 선형회귀분석에서 오차의 대칭성을 검정하기 위해 2절에서의 검정법들을 잔차에 적용한 모의실험 결과를 제시하였다. 4절에서는 회귀계수의 추정에 대해서 제안하고 여러가지 오차분포에서 모의실험을 통하여 회귀계수 추정법들을 비교하였다.

2. 검정통계량

오차의 분포가 대칭인지 또는 한쪽(오른쪽)으로 치우쳤는지는 단일표본 문제에서 미지의 중앙값

1) 이 연구는 1993년도 한국학술진흥재단 신진연구인력 연구장려금의 지원에 의한 것임
2) (151-742) 서울특별시 관악구 신림동 산56-1 서울대학교 자연과학대학 통계연구소

에 대한 분포의 대칭성 검정법들을 잔차에 적용함으로서 검정이 가능해진다. 본 연구에서는 미지의 중앙값에 대한 분포의 대칭성 검정법 중에서 Davis 와 Quade(1978), Randles et al.(1980)이 제안한 triples 검정, Boos(1982)의 검정법과 Kim(1993)의 부호순위통계량에 기초한 검정법을 오차의 대칭성 검정에 사용하여 효율성을 비교하려고 하며 이들의 검정통계량은 각각 다음과 같다. $X_1, X_2, \cdots, X_n$는 연속분포 $F(x-\theta)$에서의 확률표본이고 θ는 미지의 중앙값이라고 하자.

1) triples 검정
이 검정법은 커널(kernel)함수가

$$h(x_1, x_2, x_3) = \{sign(x_1+x_2-2x_3) + sign(x_1+x_3-2x_2) + sign(x_2+x_3-2x_1)\} / 3$$

인 U-통계량

$$\hat{\eta} = \binom{n}{3}^{-1} \sum_{i<j<k} h(x_i, x_j, x_k)$$

에 기초하며, $\hat{\eta}$의 평균과 분산은 각각

$$E[\hat{\eta}] = \eta = P\{X_1+X_2-2X_3 > 0\} - P\{X_1+X_2-2X_3 < 0\},$$
$$Var[\hat{\eta}] = \sigma_n^2 / n = \binom{n}{3}^{-1} \sum_{c=1}^{3} \binom{3}{c}\binom{n-3}{3-c}\zeta_c \qquad (2.1)$$

이고, 여기서

$$\zeta_c = Var[h_c(X_1, \ldots, X_c)]$$
$$h_c(x_1, \ldots, x_c) = E[h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_3)]$$

이다. $\hat{\eta}$이 U-통계량이므로 $\sqrt{n}(\hat{\eta}-\eta)/\sigma_n$은 점근적으로 표준정규분포를 따른다. $\hat{\zeta}_c$들은 각각 다음과 같이 추정될 수 있다.

$$\hat{\zeta}_1 = n^{-1}\sum_{i=1}^{n}(\hat{h_1}^*(X_i) - \hat{\eta})^2, \quad \hat{\zeta}_2 = \binom{n}{2}^{-1}\sum_{j<k}(\hat{h_2}^*(X_j, X_k) - \hat{\eta})^2,$$

$$\hat{\zeta}_3 = 1/9 - \hat{\eta}^2,$$

여기서

$$\widehat{h_1^*}(X_i) = \binom{n-1}{2}^{-1} \sum_{\substack{j<k \\ j\neq i\neq k}} h^*(X_i, X_j, X_k),$$

$$\widehat{h_2^*}(X_j, X_k) = (n-2)^{-1} \sum_{\substack{i=1 \\ i\neq j\neq k \\ i\neq k}} h^*(X_i, X_j, X_k)$$

이다. σ^2에 관한 식(2.1)에서 $\zeta_1, \zeta_2, \zeta_3$대신에 추정량 $\hat{\zeta}_1$, $\hat{\zeta}_2$, $\hat{\zeta}_3$을 대치함으로써 얻어지는 추정량 $\hat{\sigma}^2$은 일치추정량이다.

$\hat{\eta}$를 이용해서 검정할 수 있는 가설은

$$H_o : \eta = 0 \quad 대 \quad H_1 : \eta \neq 0 \ (또는\ \eta > 0\ ,\ \eta < 0\ )$$

이고, 비대칭분포 중에서 $\eta = 0$인 분포는 매우 드물다는 점에서 위의 귀무가설은 ''H_o : 모든 x에 대해서 $F(\theta - x) = 1 - F(\theta + x)$'' 과 동일시 해도 무리가 없다.

2) Boos의 검정법

이 방법은 Boos(1982)가 제안한 것으로 검정통계량

$$T_n = n(-1 + \sum_{i=1}^{n}\sum_{j=1}^{n} | X_i + X_j - 2\,\hat{\theta}_{HL} | \ / \ 2\sum_{i<j} | X_i - X_j |)$$

이고, 여기서 $\hat{\theta}_{HL}$은 단일표본 핫지-레만(Hodges-Lehmann)추정량으로

$$\hat{\theta}_{HL} = Med\{(X_i + X_j)/2,\ 1 \le i \le j \le n\}$$

이다. F_n을 표본분포함수라고 할 때,

$$T_n = \int_{-\infty}^{\infty} [\ F_n(\ \hat{\theta}_{HL} + x) + F_n(\ \hat{\theta}_{HL} - x) - 1]^{\,2} dx \ / \ 2n^3 \sum_{i<j} | X_i - X_j |$$

으로 나타내 진다.

T_n의 분포는 모집단 분포에 의존하고, Boos(1982)는 로지스틱분포에서 T_n분포의 백분위수를

구하여 검정에 사용할 수 있게 하였다.

3) 부호순위검정법

부호순위를 이용한 대칭성 검정통계량은 다음과 같다.

$$S^* = \sum_{i=1}^{n} \Psi_i \, a_n(R_i^*)$$

여기서 $a_n(\cdot)$은 점수(score)이고, R_i^*는 $\hat{\theta}$가 표본중앙값일 때 $|X_1 - \hat{\theta}|, \cdots, |X_n - \hat{\theta}|$ 중에 $|X_i - \hat{\theta}|$ 의 순위이며,

$$\Psi_i = \begin{cases} 1, & \text{if } X_i - \hat{\theta} \geq 0 \\ 0, & \text{if } X_i - \hat{\theta} < 0 \end{cases}$$

이다. Kim(1993)은 점수로 $a_n(i) = i^2$를 사용한 부호순위통계량

$$S^* = \sum_{i=1}^{n} \Psi_i \, R_i^{*2}$$

을 제안하였고, $n^{-1/2}(n+1)^{-2}(S^* - n(n+1)(2n+1)/12) / \sigma_s$ 이 점근적으로 표준정규분포를 따름을 보였다. $f(x)$가 X_i의 확률밀도함수일 때,

$$\sigma_s^2 = 1/45 + (b_s/f(0) - 2/3)^2 / 16$$

이고

$$b_s = 8\int_0^{\infty} (2F(x)-1) f^2(x) \, dx$$

이다.

위에서 보듯이 σ_s^2는 분포무관이 아니며, Kim(1993)은 Boos와 마찬가지로 로지스틱분포에서의 σ_s^2을 구하여 검정에 사용하였다.

3. 오차의 대칭성 검정에 대한 모의실험

이 절에서는 단순 직선 회귀모형에서 오차의 분포가 대칭인지 오른쪽으로 치우쳤는지를 검정하는 모의실험 결과를 요약한다. 2절에서의 검정법들을 잔차에 적용하였고 검정에 쓰인 잔차는 최소제곱잔차, Huber의 M-추정법에 의한 잔차 (Huber,1981), 그리고 L_1-추정법에 의한 잔차이다.

본 실험에서 Huber의 M-추정량 $\hat{\beta}_0$, $\hat{\beta}_1$은 다음의 연립방정식

$$\begin{cases} \sum_{i=1}^{n} \psi(e_i/\hat{\sigma}) = 0 \\ \sum_{i=1}^{n} \psi(e_i/\hat{\sigma}) x_i = 0 \\ \sum_{i=1}^{n} \psi^2(e_i/\hat{\sigma})/n = 0.516 \end{cases}$$

의 해이고, 여기서

$$\psi(x) = \max\{-1, \min(1, x)\}$$

이고,

$$e_i = Y - \hat{\beta}_0 - \hat{\beta}_1$$

이며 $\hat{\sigma}$는 크기모수의 보조추정량(auxiliary estimator)이다.

실험은 Kappenman(1988a,b)의 모의실험을 기초로 설계하였으며 사용된 모형은

$$Y_i = \beta_o + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

이고, 설계점은

$$x_i = w_i / \sqrt{\sum w_i^2}\ , \text{ 여기서 } w_i = \Phi^{-1}\{i/(n+1)\},$$

이다. 회귀계수들은 일반성을 잃지 않고 각각 $\beta_o = 0$, $\beta_1 = 1$로 할 수 있다. 오차의 분포는 대칭분포로 정규분포, t-분포(자유도 3,6), 로지스틱분포, 이중지수분포, 코쉬분포를 포함시켰다. 한쪽으로 치우친 분포로는 Ramberg 와 Schmeiser(1974)에 의해 논의된 일반화 람다군(the generalized lambda family, GLF) 중에서 8개를 선택하였다. 이 분포군은 누적분포함수의 역함수에 의해 정의되어 있으며 다음식

$$F^{-1}(u) = \lambda_1 + [\ u^{\lambda_3} - (1-u)^{\lambda_4}]\ /\lambda_2\ , \quad 0 < u < 1$$

에 의해 쉽게 생성될 수 있다. 모의실험에 사용된 8개 GLF분포의 모수들과 각 분포의 왜도, 첨도는 표3.1에 정리되어 있다.

표 3.1 모의실험에 사용된 GLF분포의 모수와 왜도 및 첨도

분포	λ_1	λ_2	λ_3	λ_4	왜도	첨도
GLF1	0.0	1.0	1.4	0.25	0.5	2.2
GLF2	0.0	1.0	0.00007	0.1	1.5	5.8
GLF3	3.586508	0.04306	0.025213	0.094029	0.9	4.2
GLF4	0.0	-1.0	-0.0075	-0.03	1.5	7.5
GLF5	-0.116734	-0.351663	-0.13	-0.16	0.8	11.4
GLF6	0.0	-1.0	-0.1	-0.18	2.0	21.2
GLF7	0.0	-1.0	-0.001	-0.13	3.16	23.8
GLF8	0.0	-1.0	-0.0001	-0.17	3.88	40.7

표본크기는 n=30으로 하였고, 1000번 반복 실험하여 유의수준 5%에서 귀무가설이 기각되는 횟수를 세어 표3.2를 얻었다. 대칭인 분포에서의 값은 경험적 유의수준이 되고, 오른쪽으로 치우친 분포에서의 값은 경험적 검정력을 나타낸다.

표3.2 경험적 유의수준과 검정력 ($\alpha = 0.05$)

검정법	Triples 검정			Boos 검정			부호순위검정		
분포 \ 잔차	LSE	HM	L_1	LSE	HM	L_1	LSE	HM	L_1
정규분포	.043	.047	.049	.039	.048	.051	.066*	.060	.062
t_6-분포	.049	.045	.046	.066*	.071*	.080*	.042	.049	.044
t_3-분포	.066*	.060	.066*	.069*	.068*	.076*	.047	.050	.050
로지스틱분포	.052	.055	.055	.047	.055	.056	.043	.041	.053
이중지수분포	.059	.061	.061	.070*	.076*	.085*	.046	.042	.044
코쉬분포	.059	.062	.054	.198*	.314*	.321*	.033*	.033*	.036*
GLF1	.419	.428	.430	.316	.321	.312	.398	.386	.352
GLF2	.864	.878	.853	.725	.761	.753	.566	.582	.531
GLF3	.373	.385	.390	.222	.236	.248	.228	.228	.227
GLF4	.580	.587	.572	.391	.414	.429	.307	.306	.302
GLF5	.105	.106	.106	.072	.083	.090	.065	.063	.065
GLF6	.256	.255	.250	.180	.190	.210	.148	.146	.145
GLF7	.915	.948	.929	.820	.890	.866	.611	.676	.655
GLF8	.909	.959	.938	.829	.896	.891	.594	.659	.653

0.05 근방에서 검정력의 2×표준오차는 0.014로 *는 이 범위 밖의 값을 나타냄

표3.2에서 분포무관인 Triples 검정의 유의수준이 매우 잘 만족되고 있음을 확인할 수 있고, 검정력에서도 다른 검정에 비해서 우수함을 알 수 있다. Boos의 검정은 t-분포,이중지수분포, 코쉬분포와 같이 꼬리가 두터운 분포에서 유의수준을 만족시키지 못하고 모두 2×표준오차 범위를 벗어나고 있다. Kim의 부호순위에 기초한 검정은 다소 낮은(conservative) 경험적 유의수준을 보이면서 비교적 안정된 수준을 유지하고 있으나 검정력이 triples 검정보다 떨어지는 것을 알 수 있다. 잔차중에서는 M-추정량 잔차(HL)를 이용했을 때 다른 잔차를 이용했을 때보다 검정력이 높게 나타났지만 그 차이는 검정력에 의한 차이와 비교할 때 크지 않아서 잔차보다는 검정법이 검정력의 우열을 결정하는 것으로 나타났다.

오차분포의 대칭성 검정에서 위의 모의실험 결과에서 나타난 Huber의 M-추정량 잔차를 이용한 triples 검정의 비교우위성을 토대로 다음절에서는 이러한 검정을 이용한 회귀계수 추정에 대해서 생각해 보고자 한다.

4. 회귀계수의 추정

오차의 분포가 한쪽(오른쪽)으로 치우쳐 있을 때의 회귀계수 추정은 Kappenman(1988a)이 제안한 추정법을 적용시킬 수 있다. Kappenman이 제안한 회귀계수의 추정량 $\hat{\beta}_0$, $\hat{\beta}_1$은

$$\sum \frac{x_i - \bar{x}}{y_i - q(\hat{\beta}_1) - \hat{\beta}_1 x_i} = 0 \quad \text{와} \quad \hat{\beta}_0 = median\{y_i - \hat{\beta}_1 x_i\} \tag{4.1}$$

의 해로 주어지고, 여기서 $q(\hat{\beta}_1) = \sum c_i z_i$이고, $z_1, \ldots, z_n$은 $y_1 - \hat{\beta}_1 x_1, \ldots, y_n - \hat{\beta}_1 x_n$의 순서통계량이며,

$$c_1 = 1 + (1-1/n)^n, \; c_i = 1 + (1-i/n)^n - (1-(i-1)/n)^n, \; i = 2, \ldots, n$$

이다. 위 (4.1)의 왼쪽식은 오차의 모분포를 감마분포로 하여 최우추정법과 Cooke(1979)의 위치모수에 대한 비모수적 추정법을 이용하여 얻어진 것이다.

오차의 분포가 오른쪽으로 치우쳐 있을 때는 위와같이 추정하는 것이 분포의 대칭성을 가정하고 있는 최소제곱추정이나 M-추정량같은 로버스트추정보다 좋은 추정이 될 것이라고 기대할 수 있다. 따라서 본 연구에서는 선형회귀분석에서 오차분포에 따른 적응(adaptive) 회귀계수 추정법을 제안하며 이는 회귀계수를 추정할 때 먼저 오차분포의 대칭성에 대해서 Huber의 M-추정잔차를 사용한 Triples 검정을 하고 그 결과 치우친 분포일 경우에는 (4.1)의 해로 회귀계수를 추정하고, 대칭 분포일 때는 M-추정법을 사용하는 것이다.

이 절에서 제안한 회귀계수 추정방법과 다른 추정법들을 비교하기 위해서 모의실험을 하였다.

실험설계는 3절과 같고 추정 방법으로는 최고제곱추정법(LSE), Huber의 M-추정법(HM), L_1-추정법(L1)과 본 연구에서 제안한 적응 추정법(AM)을 포함시켰다. 회귀계수의 추정에서 주요관심은 기울기의 추정에 있으므로 모의실험을 통하여 기울기 추정량의 효율성을 비교 하였다. 그 결과를 표4.1에 요약하였는데 이 표의 값들은 추정법들의 평균제곱오차(mean squared error) 중에서 최소값을 각 추정법의 평균제곱오차로 나누어서 얻어진 것이다. 따라서 표4.1에서는 추정법중에서 가장 작은 평균제곱오차를 가지면 1로 표시된다.

표에서 회귀계수의 최소제곱추정법이 꼬리가 두터운 분포나 비대칭분포에서 상대적으로 효율이 떨어짐을 확인할 수 있다. L_1-추정법과 Huber의 M-추정법은 대칭인 분포에서 효율적이고 로버스트한 특성을 보이는 반면 비대칭분포에서 L_1-추정법은 0.152에서 0.831사이, Huber의 M-추정법은 0.231에서 1.000사이의 값으로 나타나 제안된 적응 추정법과 비교할 때 효율이 떨어진다. 이들과 비교할 때, 제안된 적응추정법은 비대칭분포에서 .874에서 1.000사이의 값으로 높은 효율을 가질 뿐만아니라 대칭분포에서의 효율도 .651에서 .962사이의 높은 수준을 유지하는 것으로 나타났다.

전체적으로 본 연구에서 제안한 적응추정법은 대칭, 비대칭분포를 포함한 넓은 범위의 분포에서 효율성이 우수하고 robust한 것으로 나타났다.

표4.1 기울기 추정량들의 효율 : 가장 작은 평균제곱오차에 대한 상대비

분포 \ 추정법	LSE	L_1	HM	AM
정규분포	1.000	.633	.943	.893
t_6-분포	.893	.762	1.000	.962
t_3-분포	.587	.820	1.000	.939
로지스틱분포	.977	.707	1.000	.953
이중지수분포	.723	.986	1.000	.952
코쉬분포	.000	1.000	.824	.651
GLF1	.957	.399	.778	1.000
GLF2	.309	.236	.412	1.000
GLF3	.966	.656	1.000	.937
GLF4	.846	.733	1.000	.917
GLF5	.709	.831	1.000	.916
GLF6	.627	.763	1.000	.874
GLF7	.120	.167	.257	1.000
GLF8	.091	.152	.231	1.000

5. 결론

분포의 대칭성 검정에서는 비대칭분포의 광범위성 때문에 검정법들의 우열을 이론적으로 밝히기 어렵다. 그러나 모의실험을 통해서 검정법들의 우열을 가늠해 볼 수 있으며 본 연구의 결과도 거기서 의미를 찾을 수 있으리라고 본다. 선형회귀모형에서 오차 분포의 대칭성 검정이 실제 문제에서 분석의 목표가 되는 경우는 많지 않을 것이다. 그 보다는 다음 단계의 분석을 위한 기초 분석과정으로써 더 의미를 가질 것이며 그런 과정에서 본 연구의 결과가 참고되기를 기대 한다.

참 고 문 헌

[1] Boos, D.D (1982). A Test for Asymmetry Associated With the Hodges-Lehmann Estimator, *Journal of the American Statistical Association,* Vol. 77, 647-651.

[2] Cooke, P. (1979). Statistical inference for bounds of random variables, *Biometrika*, Vol. 66, 367-374.

[3] Davis, C. E. and Quade, D. (1978). U-statistics for Skewness or Symmetry, *Communications in Statistics, Part A-Theory and Methods*, vol. 7, 413-418.

[4] Huber, P.J. (1981). *Robust Statistics*, John Wiley, New York.

[5] Kappenmam, R.F (1988a). Robust symmetric distribution location estimation and regression, *Journal of Statistical Planning and Inference*, Vol. 19, 55-72.

[6] Kappenmam, R.F (1988b). Detecting of Symmetry or Lack of It and Applications, *Communications in Statistics-Theory and Methods*, Vol. 17(12), 4163-4177.

[7] Kim, S.O (1993). Tests for Asymmetry Associated with the Linear Signed Rank Statistics, 「한국품질학회지」, 제21권 1호, 136-143.

[8] Ramberg, J.S. and Schmeiser, B.W (1974). An Approximate Method for Generating Asymmetric Random Variables, *Communications of the ACM*, 17, 78-82.

[9] Randles, R.H, Fligner, M.A, Policello, G.E and Wolfe, D.A (1980). An Asymptotically Distribution-Free Test for Symmetry versus Asymmetry, *Journal of the American Statistical Association*, Vol. 75, 168-172.