응용통계연구 제 17권 2호, 2004년, pp. 269-280

순위회귀모형의 새로운 스코어 함수의 효율성 연구 *

최영훈 [1)]

요 약

본 연구를 통하여 순위를 이용한 선형회귀모형의 가정된 분포형태가 우리가 실질적으로 많이 접하게 되는 비대칭분포이면서 만일 가정된 분포가 오른쪽으로 늘어진 경우일 때에는, 본 논문에서 제안된 스코어의 가능한 한 0 보다 크고 1 보다 작은 r 및 1 보다 큰 s 를 선택 $(0 < r < 1, s > 1)$ 하는 것이 윌콕슨 스코어보다 높은 효율성을 나타낸다. 이와 반대로 만일 가정된 비대칭분포가 왼쪽으로 늘어진 경우일 때에는, 제안된 스코어의 가능한 한 1 보다 큰 r 및 0 보다 크고 1 보다 작은 s 를 선택 $(r > 1, 0 < s < 1)$ 하는 것이 윌콕슨 스코어보다 높은 효율성을 나타낸다. 아울러 바람직한 r 과 s 를 결정하기 위한 가정된 분포의 대칭성 검정기법도 제시한다.

주요용어: 주요용어: 선형회귀모형, 스코어, 산포함수, 점근 상대효율(ARE), 일반화된 F 분포.

1. 서론

최근들어 순위를 이용한 선형회귀모형에 대한 추정량 연구에 상당한 관심과 발전을 이루어왔다. Jureckova(1969, 1971) 와 Jaeckel(1972) 이 산포함수를 정의한 이래로 McKean and Sievers(1989), Hettmansperger and McKean(1998), Witt, Naranjo and McKean(1995) 등이 선형회귀모형에서의 순위를 이용한 영향함수, 파손 추정량 등의 이론적 근거를 제시하였다.

Ozturk and Hettmansperger(1996) 및 Choi(1998)는 위치 및 척도모수에 대한 로버스트 추정량을 유도하였으며, Ahmad(1996)는 가정된 분포의 왼쪽 꼬리부분의 제곱승을 고려한 일련의 Mann-Whitney-Wilcoxon 유형의 검정통계량을 발전시켰다. 반면에 Ozturk and Hettmansperger(1997) 및 Ozturk(1999, 2001)는 만일 이상치 자료의 정보를 모르거나 이상치의 표본자료가 존재한다면 선형회귀모형의 추정량은 자료분포의 오른쪽과 왼쪽 꼬리부분을 모두 반영하여야만 로버스트함을 주장하고 있다. 따라서 Choi and Ozturk(2002)는 선형회귀모형의 가정된 분포의 오른쪽과 왼쪽 꼬리부분을 모두 반영하는 순위추정량을 위한 새로운 스코어 발생함수와 이론적 결과를 제안하였다.

* 본 연구는 2004년 한신대학교 학술연구비 지원에 의한 것임.

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과 교수
E-mail: choicyh@hanshin.ac.kr

본 연구의 주된 목적은 Choi and Ozturk(2002)가 제시한 r 및 s 거듭제곱을 포함한 순위회귀모형의 추정량을 위한 스코어 함수가 모집단 분포의 대칭 및 비대칭 유형에 따라 최적의 효율성을 보장하는 r 및 s 를 S-PLUS 통계프로그램을 이용하여 보다 구체적으로 밝혀내고자 한다.

2장에서는 새롭게 제안한 스코어 함수를 정의하고, 3장에서는 윌콕슨 스코어와 Choi and Ozturk(2002)가 제안한 스코어에 기초를 둔 선형회귀모형의 순위추정량의 효율성 비교를 모집단 분포의 유형[대칭분포, 비대칭분포(오른쪽 및 왼쪽으로 늘어진 분포), 일반화된 F 분포 등]에 따라 시도하고, 유형에 따른 가장 적절한 r 및 s 를 발견하고자 한다. 4장에서는 실제적으로 적절한 r 과 s 의 선택을 위하여 선형회귀모형의 가정된 분포의 대칭성을 검정하기 위한 기법을 제시하고자 한다.

2. 스코어 발생함수

선형회귀모형 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n,$ 단 $\boldsymbol{x}_i$ 및 $\boldsymbol{\beta}$ 는 각각 설명변수의 $p \times 1$ 벡타와 미지의 회귀모수를 나타낸다고 정의하자. ϵ_i 는 확률밀도함수 f 와 누적분포함수 F 를 갖는다. 그렇다면 우리의 관심은 회귀모수 $\boldsymbol{\beta}$ 의 순위추정이며, Jaeckel(1972) 의 일반 순위산포함수는 아래와 같이 정의한다.

$$D(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\, a[R(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})],$$

단 $a(i) = \phi(i/(n+1))$, 스코어 발생함수 $\phi(u)$ 는 $(0,1)$ 에서 $\int_0^1 \phi(u)du = 0$ 와 $\int_0^1 \phi^2(u)du = 1$ 의 조건을 만족한다.

이제 순위를 이용한 회귀추정량의 효율을 향상시키기 위하여 새로운 스코어 발생함수를 소개하고자 한다. 즉

$$\begin{aligned}
\phi(u) &= \frac{1}{\sqrt{\omega_{r,s}}}\left[u^r - \frac{1}{r+1} - (1-u)^s + \frac{1}{s+1}\right], \\
a(i) &= \frac{1}{\sqrt{\omega_{r,s}}}\left[\left(\frac{i}{n+1}\right)^r - \frac{1}{r+1} - \left(1 - \frac{i}{n+1}\right)^s + \frac{1}{s+1}\right], \\
\omega_{r,s} &= \frac{r^2}{(2r+1)(r+1)^2} + \frac{s^2}{(2s+1)(s+1)^2} + \frac{2}{(r+1)(s+1)} - 2\frac{\Gamma(r+1)\Gamma(s+1)}{\Gamma(r+s+2)}.
\end{aligned}$$

따라서 우리의 새로운 스코어 발생함수에 의한 산포함수는 아래와 같이 표현할 수 있으며

$$D_{r,s}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} e_i\, a\,[\,R(e_i)\,],$$

이때의 $R(e_i)$ 는 $e_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ 의 순위를 의미하며, $\hat{\boldsymbol{\beta}}$ 는 $D_{r,s}(\hat{\boldsymbol{\beta}})$ 을 최소화하는 순위추정량 $\hat{\boldsymbol{\beta}}_{r,s}$ 에 의하여 추정될 수 있다.

3. 점근 상대효율 비교

본 장에서는 2장에서 제안된 스코어 함수와 잘 알려진 윌콕슨 스코어와의 효율성 비교를 하고자 한다. 윌콕슨 스코어에 기초한 β 의 순위추정량 $\hat{\beta}_{1,1}$ 의 점근적 분산을 $v(\hat{\beta}_{1,1})$ 라 정의하고, 2장에서 제안된 추정량 $\hat{\beta}_{r,s}$의 점근적 분산을 $v(\hat{\beta}_{r,s})$라 정의하자. 그렇다면 추정량 $\hat{\beta}_{r,s}$ 의 추정량 에 $\hat{\beta}_{1,1}$대한 점근 상대효율은 다음과 같이 표현할 수 있다.

$$\begin{aligned} \mathrm{ARE}(11, rs) &= \left(\frac{|\, v(\hat{\beta}_{r,s})\,|}{|\, v(\hat{\beta}_{1,1})\,|}\right)^{1/p} = \frac{\omega_{r,s}}{\tau_{r,s}}\, 12 \left[\int f^2(x)dx\right]^2, \\ \text{단 } \tau_{r,s} &= \left(\int [\, rF^{r-1}(t) + s(1-F(t))^{s-1}\,]\, f^2(t)dt\right)^2 . \end{aligned}$$

이때 점근 상대효율 $\mathrm{ARE}(11, rs) < 1$ 은 우리의 제안된 스코어 함수의 효율성이 윌콕슨 스코어의 효율성보다 뛰어남을 의미한다. 이제 코쉬, 이중지수, 균일, 지수, 로그정규, 혼합정규, 베타, 일반화된 F, 강하게 늘어진 분포 등의 다양한 분포유형에 따른 점근 상대효율 $\mathrm{ARE}(11, rs)$ 의 변화를 상세하게 검토하고자 한다.

3.1. 대칭분포

본 절에서는 윌콕슨 스코어보다 향상된 스코어 함수의 r 과 s 의 선택을 구체적으로 탐구하고자 일차적으로 대칭형태를 갖는 분포에 대하여 $\mathrm{ARE}(11, rs)$ 의 결과를 비교정리하였다.

표3.1은 코쉬, 이중지수 및 $\epsilon = 0.3,\ 0.5,\ a = 3, 5, 7$ 인 척도이동에 따른 혼합정규분포와 같이 양쪽끝이 길게 늘어진 대칭분포의 $r, s = 1.1(1.9)0.2$ 의 선택된 수치값에 따른 본 논문에서 제안한 스코어의 윌콕슨 스코어에 대한 점근 상대효율을 나타낸다. 그리고 상대적으로 양쪽끝이 짧게 늘어진 대칭분포인 균일분포의 $r, s = 0.1(0.9)0.2$ 에 따른 점근 상대효율도 보여준다. 한편 그림3.1은 코쉬, 이중지수, $\epsilon = 0.5,\ a = 7$ 인 혼합정규분포 및 균일분포의 확률밀도함수와 r 과 s 를 각각 $0.1(5)0.1$ 로 증가시켰을 때의 우리의 제안된 스코어의 윌콕슨 스코어에 대한 점근 상대효율 $\mathrm{ARE}(11, rs)$ 의 투시도를 묘사한다.

표3.1 및 그림3.1은 코쉬분포, 이중지수분포 및 척도이동에 따른 혼합정규분포와 같이 양쪽끝이 길게 늘어진 형태를 갖는 대칭분포에 대하여 $\mathrm{ARE}(11, rs)$ 의 수치 및 투시도가 전반적으로 상당히 유사함을 나타낸다.

구체적으로 표3.1 및 그림3.1로부터 제안된 스코어를 이용한 추정량 $\hat{\beta}_{r,s}$ 이 양쪽끝이 모두 길게 늘어진 형태를 갖는 대칭분포하에서는 $1 < r,\ s < 2$ 일 때 제안된 추정량 $\hat{\beta}_{r,s}$ 의 점근적 분산 $v(\hat{\beta}_{r,s})$ 이 우리가 고려한 $r, s = 0.1(5)0.1$ 의 모든 수치중에서 윌콕슨 스코어를 이용한 추정량 $\hat{\beta}_{1,1}$ 보다 향상된 점근 상대효율을 보임을 알 수 있으며 공간의 제약으로 말미암아 모든 수치값을 표로 제시하지는 못하였다.

한편 표3.1 및 그림3.1에서 알 수 있는 바와같이 꼬리부분이 얇은 대칭분포인 균일분포하에서는 특히 $0 < r,\ s < 1$ 일 때 윌콕슨 스코어보다 제안된 스코어의 점근 상대효율이 상당히 뛰어남을 나타내며, 가능한 한 작은 $0 < r,\ s < 1$ 를 선택하여야 한다.

표 3.1: 대칭분포의 ARE(11, rs), 단 $N_\epsilon(0,a) = (1-\epsilon)N(0,1) + \epsilon N(0,a)$

분포 / r s	1.1	1.3	1.5	1.7	1.9
코쉬	.989	.978	.976	.981	.992
이중지수	.995	.990	.989	.991	.996
$N_{0.3}$ (0, 3)	.996	.992	.991	.993	.997
$N_{0.3}$ (0, 5)	.992	.984	.983	.987	.994
$N_{0.3}$ (0, 7)	.990	.980	.978	.983	.993
$N_{0.5}$ (0, 3)	.995	.990	.989	.991	.996
$N_{0.5}$ (0, 5)	.990	.980	.978	.983	.993
$N_{0.5}$ (0, 7)	.987	.973	.971	.977	.990
분포 /r s	0.1	0.3	0.5	0.7	0.9
균일	.111	.375	.644	.838	.960

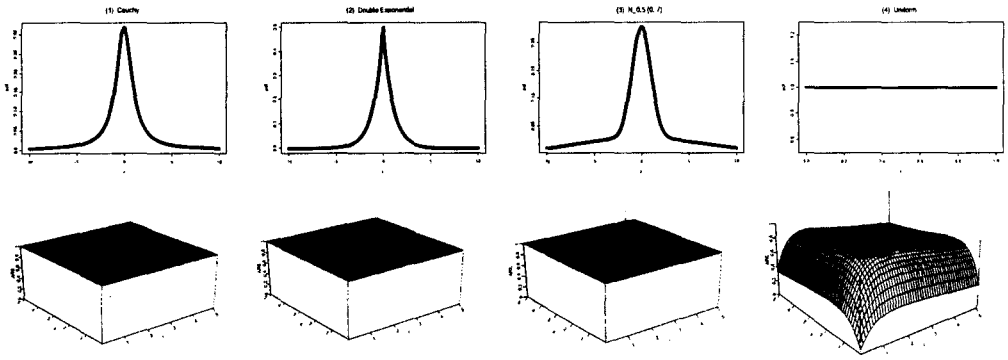

그림 3.1: 대칭분포의 확률밀도함수(pdf) 및 점근 상대효율 ARE(11, rs)의 투시도

3.2. 오른쪽으로 늘어진 비대칭분포

표3.2는 지수분포, 로그정규분포 및 $\epsilon = 0.3,\ 0.5$, $a = 3, 5, 7$, $b = 1, 2$, $r = 0.1, 0.5, 0.9, 1, 3, 5$, $s = 1, 3, 5$ 의 선택된 수치에 따른 혼합정규분포와 같이 오른쪽으로 늘어진 비대칭분포 형태에 따른 우리의 제안된 스코어의 윌콕슨 스코어에 대한 점근 상대효율 ARE(11, rs) 의 결과를 나타낸다. 이때 고려된 수치는 $r, s = 0.1(5)0.1$ 이나 선택된 결과값만 제시하였다. 한편 그림3.2는 지수분포, 로그정규분포 및 $\epsilon = 0.3,\ 0.5$, $a = 7$, $b = 1, 2$ 인 혼합정규분포의 확률밀도함수와 r 과 s 를 각각 0.1(5)0.1로 증가시켰을 때의 점근 상대효율 ARE(11, rs) 의 투시도를 묘사한다.

표 3.2: 오른쪽으로 늘어진 비대칭분포의 점근 상대효율 ARE(11, rs), 단 $N_\epsilon(a,b) = (1-\epsilon)N(0,1) + \epsilon N(a,b)$

r	0.1			0.5			0.9			1	1	3	3	5	5
분포 s	1	3	5	1	3	5	1	3	5	3	5	3	5	3	5
지수	.201	.145	.138	.600	.397	.303	.930	.570	.426	.602	.448	.814	.569	.795	.543
로그정규	.672	.405	.315	.735	.491	.400	.947	.606	.489	.631	.508	.872	.672	.884	.669
$N_{0.3}(3,1)$	.915	.795	.782	.919	.815	.786	.983	.849	.808	.858	.814	.945	.883	.920	.861
$N_{0.3}(5,1)$	.897	.723	.695	.900	.756	.718	.978	.803	.753	.814	.762	.934	.862	.900	.830
$N_{0.3}(7,1)$	.893	.705	.678	.895	.741	.703	.977	.790	.742	.803	.752	.946	.875	.935	.866
$N_{0.5}(3,2)$	.893	.722	.684	.896	.755	.710	.978	.804	.748	.816	.757	.939	.857	.927	.847
$N_{0.5}(5,2)$	.847	.630	.542	.858	.679	.596	.971	.749	.653	.765	.665	.878	.743	.820	.689
$N_{0.5}(7,2)$	.809	.530	.441	.819	.590	.506	.961	.673	.574	.693	.589	.895	.738	.897	.733

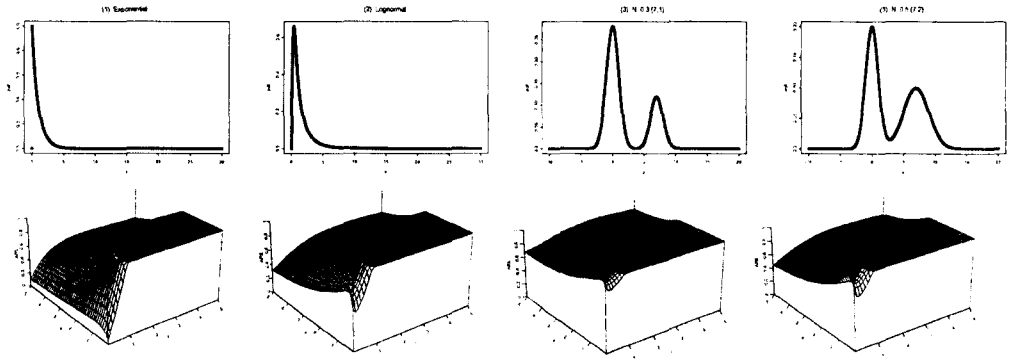

그림 3.2: 오른쪽으로 늘어진 비대칭분포의 확률밀도함수(pdf) 및 점근 상대효율 ARE(11, rs)의 투시도

표3.2 및 그림3.2의 결과는 다음과 같이 요약할 수 있다. 지수분포, 로그정규분포 및 정의된 혼합정규분포와 같이 오른쪽으로 늘어진 비대칭분포 형태에 대하여 만일 $r < 1, s > 1$ 이면 $v(\hat{\beta}_{r,s})$ 은 $v(\hat{\beta}_{1,1})$ 보다 상당히 작다. 특히 오른쪽으로 강하게 늘어진 분포에 대하여 우리의 제안된 스코어 발생함수는 r 값이 작고 s 값이 클수록 윌콕슨 스코어보다 월등히 향상된 효율성을 제공한다는 사실이다. 따라서 종합적으로 오른쪽으로 늘어진 비대칭분포의 경우에 우리는 $r = 0.1$ 과 s 는 가능한 한 큰 값을 선택하여야 한다.

3.3. 왼쪽으로 늘어진 비대칭분포

표3.3은 $\epsilon = 0.3, 0.5$, $a = -3, -5, -7$, $b = 1, 2$, $r = 1, 3, 5$, $s = 0.1, 0.5, 0.9, 1, 3, 5$ 의 선택된 수치에 따른 혼합정규분포 및 베타분포(5, 1.5) 와 같이 왼쪽으로 늘어진 비대칭분포 형태에 따른 우리의 제안된 스코어의 윌콕슨 스코어에 대한 점근 상대효율 ARE$(11, rs)$ 의 결과를 나타낸다. 이때도 고려된 수치는 $r, s = 0.1(5)0.1$ 이나 선택된 결과값만 제시하였다. 한편 그림3.3은 $\epsilon = 0.3, 0.5$, $a = -7$, $b = 1, 2$ 인 혼합정규분포 및 베타분포(5, 1.5)의 확률밀도함수와 r 과 s 를 각각 0.1(5)0.1 로 증가시켰을 때의 점근 상대효율 ARE$(11, rs)$ 의 투시도를 묘사한다.

표 3.3: 왼쪽으로 늘어진 비대칭분포의 점근 상대효율 ARE$(11, rs)$, 단 $N_\epsilon(a, b) = (1 - \epsilon)N(0, 1) + \epsilon N(a, b)$

r	1	3	5	1	3	5	1	3	5	3	5	3	5	3	5
분포 s		0.1			0.5			0.9		1	1	3	3	5	5
$N_{0.3}(-3, 1)$	.915	.795	.782	.919	.815	.786	.983	.849	.808	.858	.814	.945	.883	.920	.861
$N_{0.3}(-5, 1)$	.897	.723	.695	.900	.756	.718	.978	.803	.753	.814	.762	.934	.862	.900	.830
$N_{0.3}(-7, 1)$	.893	.705	.678	.895	.741	.703	.977	.790	.742	.803	.752	.946	.875	.935	.866
$N_{0.5}(-3, 2)$	.893	.722	.684	.896	.755	.710	.978	.804	.748	.816	.757	.939	.857	.927	.847
$N_{0.5}(-5, 2)$	.847	.630	.542	.858	.679	.596	.971	.749	.653	.765	.665	.878	.743	.820	.689
$N_{0.5}(-7, 2)$	.809	.530	.441	.819	.590	.506	.961	.673	.574	.693	.589	.895	.738	.897	.733
베타$(5, 1.5)$	.719	.558	.475	.815	.659	.574	.968	.760	.657	.779	.672	.902	.757	.889	.741

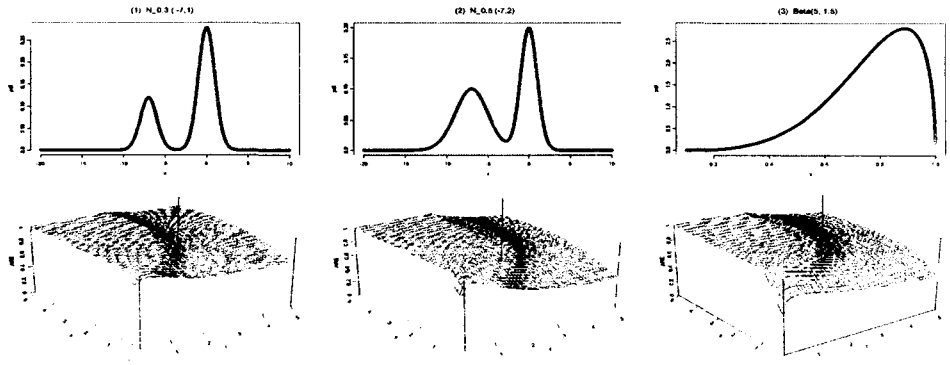

그림 3.3: 왼쪽으로 늘어진 비대칭분포의 확률밀도함수(pdf) 및 점근 상대효율 ARE$(11, rs)$의 투시도

표3.3및 그림3.3의 결과로부터 왼쪽으로 늘어진 분포형태에 대하여는 오른쪽으로 늘어진 분포형태와 상대적이며 비슷한 결과가 유도될 수 있다. 정의된 혼합정규분포 및 베타분포$(5, 1.5)$ 와 같이 왼쪽으로 늘어진 비대칭분포 형태에 대하여 만일 $r > 1$, $s < 1$ 이면 $v(\hat{\beta}_{r,s})$ 은 $v(\hat{\beta}_{1,1})$ 보다 상당히 작다. 특히 왼쪽으로 강하게 늘어진 분포에 대하여 우리의 제안된 스코어 발생함수는 r 값이 크고 s 값이 작을수록 윌콕슨 스코어보다 월등히 향상된 효율성을 제공한다는 사실이다. 요약한다면 강하게 왼쪽으로 늘어진 비대칭분포의 경우에 우리는 $s = 0.1$ 과 r 은 가능한 한 큰 값을 선택하여야 한다.

3.4. 일반화된 F 분포

이제 본 절에서는 우리의 제안된 스코어의 윌콕슨 스코어에 대한 점근 상대효율을 잘 알려진 McKean and Sievers(1989)이 제시한 일반화된 F 분포에 적용하여 비교 평가하고자 한다. 우선 F 를 자유도 $2m_1$ 과 $2m_2$ 를 갖는 F 분포의 확률변수라 정의하자. 그렇다면 $T = log(F)$ 는 자유도 $2m_1$ 과 $2m_2$ 를 갖는 일반화된 F 분포 $(GF(2m_1, 2m_2))$ 를 따른다고 한다. 일반화된 F 분포는 다양한 형태와 꼬리모양을 갖는 매우 유연한 분포로 알려져 있다. 만일 $m_1 = m_2$ 이면 대칭의 분포형태를 갖는다. 또한 만일 $m_1 > m_2$ 이면 오른쪽으로 늘어진 비대칭분포 형태를 갖고, $m_1 < m_2$ 이면 왼쪽으로 늘어진 비대칭분포 형태를 갖는다. 그리고 $m_1, m_2 < 1$ 이면 양쪽끝이 모두 길게 늘어진 대칭분포 형태를 갖고, $m_1, m_2 > 1$ 이면 양쪽끝이 모두 얇은 대칭분포 형태를 갖는다.

아래의 그림3.4는 우리의 관심대상인 한쪽끝이 길게 늘어진 일반화된 F 분포인 비대칭분포의 확률밀도함수 및 $r, s = 0.1(5)0.1$ 의 수치 변화에 따른 본 논문에서 제안한 스코어의 윌콕슨 스코어에 대한 점근 상대효율 ARE$(11, rs)$ 의 투시도를 나타낸다.

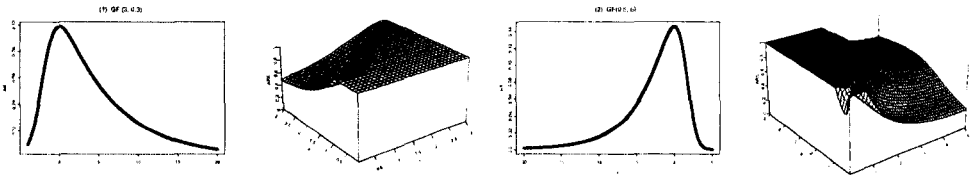

그림 3.4: 일반화된 F 분포의 확률밀도함수(pdf) 및 점근 상대효율 ARE$(11,\ rs)$ 의 투시도

위의 결과로부터 $GF(3, 0.3)$은 오른쪽으로 늘어진 분포형태를 갖고 우리의 제안된 스코어 발생함수는 r 값이 작고 s 값이 클때 윌콕슨 스코어보다 향상된 효율성을 보임을 알 수 있다. 이와 반대로 $GF(0.5, 6)$은 왼쪽으로 늘어진 분포형태를 갖고 우리의 제안된 스코어 발생함수는 r 값이 크고 s 값이 작을때 윌콕슨 스코어보다 향상된 효율성을 보임을 요약적으로 말해주고 있다.

3.5. 강하게 늘어진 분포

이외에도 위의 결과를 Xie and Priebe(2000)이 제시한 분포함수에 적용하고자 한다. 대표적으로 오른쪽으로 약하게 늘어진 단봉분포 및 오른쪽으로 강하게 늘어진 분포의 확률밀도함수는 각각

$$f(x) = \frac{1}{5}N(0,1) + \frac{1}{5}N\left(-\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(-\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$$

$$f(x) = \sum_{l=0}^{7} \frac{1}{8} N\left(3\left(\left(\frac{2}{3}\right)^l - 1\right), \left(\frac{2}{3}\right)^{2l}\right)$$

의 정규분포 결합형태로 정의되며, 이와 상대적으로 왼쪽으로 약하게 늘어진 단봉분포 및 왼쪽으로 강하게 늘어진 분포의 확률밀도함수는 각각

$$f(x) = \frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$$

$$f(x) = \sum_{l=0}^{7} \frac{1}{8} N\left(-3\left(\left(\frac{2}{3}\right)^l - 1\right), \left(\frac{2}{3}\right)^{2l}\right)$$

의 정규분포 결합형태로 정의된다.

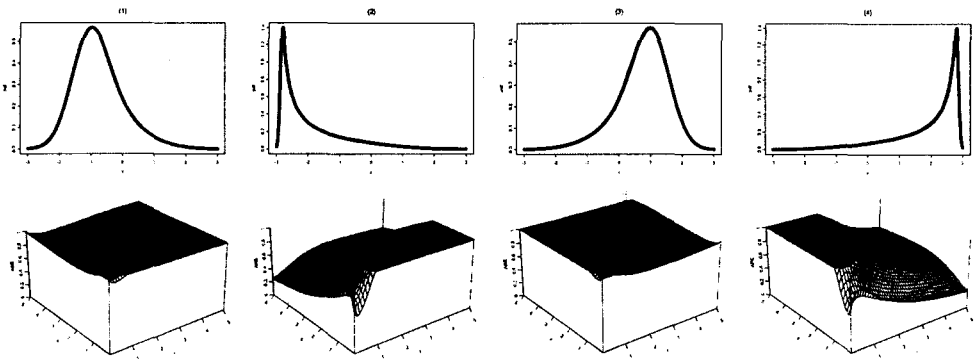

그림 3.5: 비대칭분포의 확률밀도함수(pdf) 및 점근 상대효율 ARE$(11, rs)$의 투시도

그림3.5는 이와같은 비대칭분포의 확률밀도함수 및 $r, s = 0.1(5)0.1$ 의 수치 변화에 따른 본 논문에서 제안한 스코어의 윌콕슨 스코어에 대한 점근 상대효율 ARE$(11, rs)$ 의 투시도를 대조적으로 나타낸다.

위의 결과는 다음과 같이 정리할 수 있다. 오른쪽으로 약하게 치우친 분포[(1)]는 r 이 작고 s 가 클때 ARE 값이 작아지기 시작함을 보여주며, 오른쪽으로 강하게 치우친 분포[(2)]일 때 확연히 우리의 제안된 스코어 발생함수는 윌콕슨 스코어보다 향상된 효율성을 보여준다. 한편 왼쪽으로 약하게 치우친 분포[(3)]는 r 이 크고 s 가 작을때 ARE 값이 작아지기 시

작함을 보여주며, 왼쪽으로 강하게 치우친 분포[(4)]일 때 명백히 우리의 스코어 발생함수는 윌콕슨 스코어보다 향상된 효율성을 보여준다.

4. 대칭성 검정

실제적으로 제안된 스코어 함수의 적절한 r 과 s 의 선택을 위하여 가정된 분포의 대칭성을 검정하기 위한 기법을 제시하고자 한다. 우선 가정된 확률분포 모형의 비대칭성 여부를 평가하기 위한 잔차를 구하기 위하여 윌콕슨 스코어를 사용할 것을 권장한다. 윌콕슨 잔차를 얻은 후에는, 분포의 유형이 오른쪽 혹은 왼쪽으로 늘어졌는지를 알기 위하여 대칭의 검정기법을 사용할 수 있겠다. 이를 위하여 아래에 제시된 Hollander and Wolfe(1999) 의 연속인 세 잔차 관측값에 기초한 통계량을 추천하고자 한다.

$$T = \sum_{1\le i<j<k\le n}\sum\sum f(e_i, e_j, e_k)$$

단 $f(e_i, e_j, e_k) = [sign(e_i+e_j-2e_k)+sign(e_i+e_k-2e_j)+sign(e_j+e_k-2e_i)$ 및 $t <, =, > 0$ 일 때 $sign(t) = -1, 0, 1$.

분포대칭의 귀무가설하에서 검정통계량 T 의 표준화된 수치의 값 $V = T/\hat{\alpha}$ 은 정규분포로 수렴한다. 단 $\hat{\alpha}$ 은 Hollander and Wolfe(1999) 의 공식 3.78 로 다음과 같이 주어진다.

$$\begin{aligned}\hat{\sigma}^2 &= \left[\frac{(n-3)(n-4)}{(n-1)(n-2)}\sum_{t=1}^{n} B_t^2 + \frac{(n-3)}{(n-4)}\sum_{s=1}^{n-1}\sum_{t=s+1}^{n} B_{s,t}^2 + \frac{n(n-1)(n-2)}{6}\right.\\ &- \left.\left\{1 - \frac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)}\right\}T^2\right],\end{aligned}$$

단 $t = 1, 2, \ldots, n$ 에 대하여

$$B_t = \sum_{j=t+1}^{n-1}\sum_{k=j+1}^{n} f(e_t, e_j, e_k) + \sum_{j=1}^{t-1}\sum_{k=t+1}^{n} f(e_j, e_t, e_k) + \sum_{j=1}^{t-2}\sum_{k=j+1}^{t-1} f(e_j, e_k, e_t),$$

및 $1 \le s < t \le n$ 에 대하여

$$B_{s,t} = \sum_{j=1}^{s-1} f(e_j, e_s, e_t) + \sum_{j=s+1}^{t-1} f(e_s, e_j, e_t) + \sum_{j=t+1}^{n} f(e_s, e_t, e_j).$$

그러므로 만일 $V \ge z_\alpha$ 인 경우에는 오른쪽으로 늘어진 분포형태의 대립가설이 참이므로 분포대칭의 귀무가설을 기각한다. 따라서 $r = 0.1$ 과 $s = 5$ 를 선택하면 된다. 구체적으로 적절한 r 및 s 는 $r = 1 - 0.13 \mid V \mid$, $s = 1 + 0.25 \mid V \mid$ 의 관계식으로부터 계산할 수 있다. 이와 마찬가지로, 만일 $V \le -z_\alpha$인 경우에는 왼쪽으로 늘어진 분포형태의 대립가설이 참이므로 분포대칭의 귀무가설을 기각한다. 따라서 $r = 5$ 와 $s = 0.1$ 를 선택하면 된다. 구체적으로 적절한 r 및 s 는 $r = 1 + 0.25 \mid V \mid$, $s = 1 - 0.13 \mid V \mid$ 의 관계식으로부터 계산할 수 있다.

[예제] Ott(1984, p475)의 예 13.8의 다중회귀모형을 고려하여 보자. Wilcoxon 스코어를 이용한 잔차들이 Minitab 의 RREGRESS 명령문으로부터 다음과 같이 가능하다.

0.195, -1.160, 1.247, -1.826, 0.611, 0.231, 3.007, -0.246, -2.022
0.281, -2.663, 0.195, -1.509, -4.370, 1.743, 1.584, 0.246, 1.743

그렇다면 주어진 잔차로부터 위에 제시된 공식을 이용하여 $T = -136$, $\hat{\sigma}^2 = 4132$, $V = -2.116$ 을 얻을 수가 있다. 따라서 유의수준 $\alpha = 0.05$하에서 왼쪽으로 늘어진 비대칭분포형태임을 알 수 있으며, 분포대칭의 귀무가설을 확실히 기각할 수 있다. 그러므로 $r = 1.5$, $s = 0.7$ 이 적절한 선택임을 관계식으로부터 유도할 수 있다.

5. 결론

본 연구를 통하여 선형회귀모형의 가정된 분포형태가 대칭일 때에는 (i) 양쪽끝이 길게 늘어진 경우에는 $1 < r, s < 2$ 이 (ii) 양쪽끝이 얇은 경우에는 $0 < r, s < 1$ 이 바람직하다. 그리고 우리가 실질적으로 많이 접하게 되는 비대칭분포일 때에는 (iii) 만일 가정된 분포가 오른쪽으로 늘어진 경우일 때에는, 우리의 제안된 스코어의 가능한 한 0 보다 크고 1 보다 작은 r 및 1 보다 큰 s 를 선택($0 < r < 1$, $s > 1$)하는 것이 윌콕슨 스코어보다 높은 효율성을 나타낸다. 이와 반대로 (iv) 만일 가정된 분포가 왼쪽으로 늘어진 경우일 때에는, 제안된 스코어의 가능한 한 1 보다 큰 r 및 0 보다 크고 1 보다 작은 s 를 선택($r > 1$, $0 < s < 1$)하는 것이 윌콕슨 스코어보다 높은 효율성을 나타낸다. 아울러 바람직한 r 과 s 를 결정하기 위한 가정된 분포의 대칭성 검정기법도 제시하였다.

참고문헌

Ahmad, I. A. (1996). A class of Mann-Whitney-Wilcoxon type statistics, *The American Statistician*, **50**, 324-327.

Choi, Y. H. (1998). A study of the power of the rank transform test in a 2(3) factorial experiment, *Communications in Statistics*, **27**, 251-266.

Choi, Y. H. and Ozturk, O (2002). A new class of score generating functions for regression models, *Statistics & Probability Letters*, **57**, 205-214.

Hettmansperger, T. P. and McKean J. W. (1998). *Robust Nonparametric Statistical Methods*, Wiley & Jones Inc., New York.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*, John Wiley & Sons, Inc., New York.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of residuals, *The Annals of Mathematical Statistics*, **43**, 1449-1458.

Jureckova, J. (1969). Asymptotic linearity of a rank statistics in regression parameter, *The Annals of Mathematical Statistics*, **42**, 1328-1338.

Jureckova, J. (1971). Nonparametric estimate of regression coefficients, *The Annals of Mathematical Statistics*, **42**, 1328-1338.

McKean, J. W. and Sievers, G. L. (1989). Rank scores suitable for analyses of linear models under asymmetric error distributions, *Technometrics*, **31**, 207-218.

Ott, L. (1984). *An Introduction to Statistical Methods and Data Analysis*, Duxbury, Boston, Massachusetts.

Ozturk, O (1999). Two-sample inference based on one-sample ranked set sample sign statistics, *Journal of Nonparametric Statistics*, **10**, 197-212.

Ozturk, O (2001). A generalization of Ahmad's class of Mann-Whitney-Wilcoxon statistics, *Australian and New Zealand Journal of Statistics*, **43(1)**, 67-74.

Ozturk, O and Hettmansperger, T. P. (1996). Almost fully efficient and robust simultaneous estimation of location and scale parameters: a minimum distance approach, *Statistics & Probability Letters*, **29**, 233-244.

Ozturk, O and Hettmansperger, T. P. (1997). Generalized weighted Cramer-Von Mises distance estimators, *Biometrika*, **84**, 283-294.

Witt, L. D., Naranjo, J. D. and McKean, J. W. (1995). Influence functions for rank based procedures in the linear model, *Journal of Nonparametric Statistics*, **5**, 339-358.

Xie and Priebe (2000). Generalizing the Mann-Whitney-Wilcoxon statistic, *Nonparametric Statistics*, **12**, 661-682.

[2003년 9월 접수, 2003년 11월 채택]

Asymptotic Relative Efficiency for New Score Functions in Rank Regression Models *

Young Hun Choi [1)]

ABSTRACT

We explore the selection of r and s that provides improvement over the Wilcoxon scores under the asymmetric distributions we encounter in practice. We select $0 < r < 1, s > 1$ for right-skewed distribution and $r > 1, 0 < s < 1$ for left-skewed distributions from the perspective plots. We also study the association between the desirable r and s and the test statistic for skewness.

Keywords: Wilcoxon score, Rank regression model, Dispersion function, Asymptotic relative efficiency, Generalized F distribution

* This research was supported by Hanshin University Research Grant 2004.

1) Professor, Department of Information and Statistics, Hanshin University, 411 Yangsan-dong, Osan, Kyunggi-do, 447-791, Korea.
E-mail: choicyh@hanshin.ac.kr