The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 435-446

Asymptotic Relative Efficiency for New Scores in the Generalized F Distribution [1]

Young Hun Choi [2]

Abstract

In this paper we introduced a new score generating function for the rank dispersion function in a multiple linear model. Based on the new score function, we derived the asymptotic relative efficiency, ARE(11, rs), of our score function with respect to the Wilcoxon scores for the generalized F distributions which show very flexible distributions with a variety of shape and tail behaviors. We thoroughly explored the selection of r and s of our new score function that provides improvement over the Wilcoxon scores.

Keywords : Scores; Dispersion function; Asymptotic Relative Efficiency; Generalized F distribution.

1. Introduction

Recently Ozturk and Hettmansperger(1996) and Ozturk(1999) derived the robust estimates of location and scale parameters by minimizing distance criterion function. Ahmad(1996) developed a new class of Mann-Whitney-Wilcoxon type test statistics which only considered the one side tail probabilities of the underlying distribution. Ozturk and Hettmansperger(1997) considered the distribution functions reflecting both right and left tail probabilities. Ozturk(2001) considered another class of Mann-Whitney-Wilcoxon test statistics by incorporating both right and left tail behavior of the underlying distributions. Further Choi and Ozturk(2002) introduced a new score generating function for the rank dispersion function in a multiple linear regression model which improved the efficiency for many distributions by comparing the score function with the rth and sth power of the tail probabilities of the underlying probability distributions. Choi(2004) explored efficiency comparison over the

1) This research was supported by Hanshin University Research Grant in 2004.
2) Professor, Department of Information and Statistics, Hanshin University, 447-791.
E-mail : choicyh@hanshin.ac.kr

Wilcoxon scores under the asymmetric distributions in essence.

Now the main purpose of this paper is to extend the Hettmansperger and McKean(1998) and Choi and Ozturk(2002)'s concept, where the distribution function reflects on both right and left tail probabilities and produces robust estimators with high efficiency, into the rank estimate of regression parameters in a linear model. In addition, this paper is to extend Choi(2004)'s result in practice since the generalized F distributions are more realistic and applicable with respect to the diversified scheme of distributions.

In Section 2, we propose our score function based on the rth and sth power in considering both right and left tail probabilities. We define the dispersion function $D_{r,s}(\beta)$ based on our rth and sth power score function. In Section 3, we define the asymptotic relative efficiency, ARE(11, rs), of our score function with respect to the Wilcoxon scores. In Section 4, we compare the efficiency of rank estimator based on our proposed score generating function with the efficiency of rank estimator based on the Wilcoxon scores for the generalized F distributions. We thoroughly explore the selection of r and s that provides improvement over the Wilcoxon scores.

2. Score Function

Consider the linear regression model,

$$y_i = \alpha + x_i'\beta + e_i$$

for $i = 1, \cdots, n$, where x_i and β are $p \times 1$ vectors of explanatory variables and unknown regression parameters respectively and e_i is a random variable with density f and distribution function F.

Jaeckel's(1972) general rank dispersion function can be defined as

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

where a set of scores is generated by $a(i) = \phi(i/(n+1))$, and the score generating function $\phi(u)$ is a nondecreasing, square integrable, bounded function on $(0,1)$ and satisfies the conditions $\int_0^1 \phi(u)\,du = 0$ and $\int_0^1 \phi^2(u)\,du = 1$.

Now let

$$\phi(u) \;=\; \frac{1}{\sqrt{\omega_{r,s}}}\left[u^r - \frac{1}{r+1} - (1-u)^s + \frac{1}{s+1} \right],$$

$$a(i) = \frac{1}{\sqrt{\omega_{r,s}}}\left[\left(\frac{i}{n+1}\right)^r - \frac{1}{r+1} - \left(1-\frac{i}{n+1}\right)^s + \frac{1}{s+1}\right],$$

where $$\omega_{r,s} = \frac{r^2}{(2r+1)(r+1)^2} + \frac{s^2}{(2s+1)(s+1)^2} + \frac{2}{(r+1)(s+1)} - 2\frac{\Gamma(r+1)\,\Gamma(s+1)}{\Gamma(r+s+2)}.$$

Define the dispersion function

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

where $R(e_i)$ denotes the rank of $e_i = y_i - x_i'\beta$. Then β can be estimated by the rank estimator $\widehat{\beta_{r,s}}$ which minimizes the dispersion function.

3. Asymptotic Relative Efficiencies

In this section, we compare the efficiency of the proposed score function with respect to the Wilcoxon scores. The asymptotic variance of the rank estimate of β based on the Wilcoxon scores is denoted here as $v(\widehat{\beta_{1,1}})$. Then from Theorem 2 of Choi and Ozturk(2002), the asymptotic relative efficiency of our estimator $\widehat{\beta_{r,s}}$ with respect to $\widehat{\beta_{1,1}}$ is expressed as

$$\mathrm{ARE}(11, rs) = \left(\frac{|v(\widehat{\beta_{r,s}})|}{|v(\widehat{\beta_{1,1}})|}\right)^{1/p} = \frac{\omega_{r,s}}{\tau_{r,s}} 12\left[\int f^2(x)\,dx\right]^2, \tag{1}$$

where $\tau_{r,s} = \left(\int [rF^{r-1}(t) + s(1-F(t))^{s-1}]\, f^2(t)\,dt\right)^2.$

The asymptotic relative efficiencies ARE(11, rs), where ARE(11, rs) < 1 implies that the efficiency of our score function is superior to that of the Wilcoxon scores, are discussed below for several generalized F distributions. Let F be a random variable having an $F_{2m_1, 2m_2}$ distribution with degrees of freedoms $2m_1$ and $2m_2$. Then $T = \log(F)$ is said to have the generalized F distribution [$GF(2m_1, 2m_2)$] with degrees of freedoms $2m_1$ and $2m_2$.

Lemma 1. The generalized F distribution, $GF(2m_1, 2m_2)$, with degrees of freedoms $2m_1$ and $2m_2$ has the following relationship with the ordinary $F_{2m_1, 2m_2}$ distribution with degrees of freedoms $2m_1$ and $2m_2$.

$$GF(2m_1, 2m_2) = F_{2m_1, 2m_2}(e^t)$$

Proof. First consider a random variable F having the ordinary $F_{m,n}$ distribution with m and n degrees of freedoms. Then the probability density function can be written as

$$f_{m,n}(F) = \frac{\Gamma((m+n)/2)(m/n)^{m/2}}{\Gamma(m/2)\Gamma(n/2)} \cdot \frac{F^{m/2-1}}{(1+(m/n)F)^{(m+n)/2}} . \tag{2}$$

Let $x = \log(F)$, $m = 2m_1$ and $n = 2m_2$. By using integration by parts and substituting $F = e^x$ and $dF = e^x dx$ into (2), we can easily show that

$$f_{2m_1, 2m_2}(x) = \frac{\Gamma(m_1+m_2)(m_1/m_2)^{m_1}}{\Gamma(m_1)\Gamma(m_2)} \cdot \frac{e^{x m_1}}{[1+(m_1/m_2)e^x]^{m_1+m_2}} .$$

Therefore we can derive that $GF(2m_1, 2m_2) = F_{2m_1, 2m_2}(e^t)$.

Lemma 2. The asymptotic relative efficiency of our estimator with respect to the Wilcoxon scores for the generalized F distribution with degrees of freedoms $2m_1$ and $2m_2$ is expressed as

$$\mathrm{ARE}(11, rs) = \frac{\omega_{r,s}}{\tau_{r,s}} \frac{12\,\Gamma^4(m_1+m_2)\Gamma^2(2m_1)\Gamma^2(2m_2)}{\Gamma^4(m_1)\Gamma^4(m_2)\Gamma^2(2m_1+2m_2)} .$$

Proof. By using Lemma 1, it can be proven that for the Wilcoxon scores at the generalized F distribution

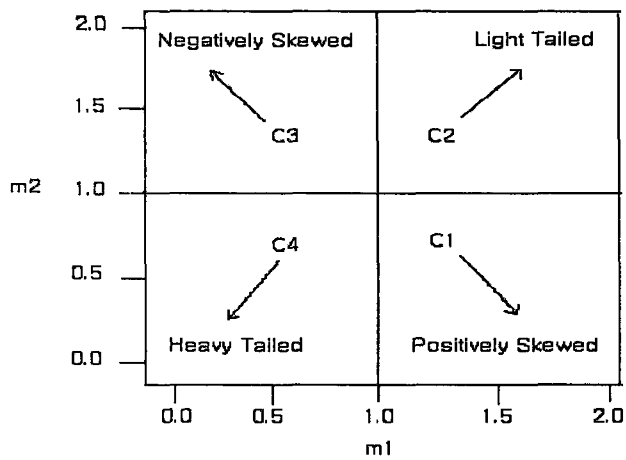
$$\begin{aligned}
&\sqrt{12}\int f^2(x)\,dx \\
&= \sqrt{12}\int \left[\frac{\Gamma(m_1+m_2)(m_1/m_2)^{m_1}}{\Gamma(m_1)\Gamma(m_2)}\right]^2 \cdot \frac{e^{x2m_1}}{[1+(2m_1/2m_2)e^x]^{2m_1+2m_2}}\,dx \\
&= \sqrt{12}\left[\frac{\Gamma(m_1+m_2)(m_1/m_2)^{m_1}}{\Gamma(m_1)\Gamma(m_2)}\right]^2 \frac{\Gamma(2m_1)\Gamma(2m_2)}{\Gamma(2m_1+2m_2)(m_1/m_2)^{2m_1}} \\
&= \frac{\sqrt{12}\,\Gamma^2(m_1+m_2)\Gamma(2m_1)\Gamma(2m_2)}{\Gamma^2(m_1)\Gamma^2(m_2)\Gamma(2m_1+2m_2)} .
\end{aligned}$$

Therefore when substituting the above result into (1), ARE(11, rs) for the generalized F distribution can be obtained straightforwardly.

4. Efficiency Comparisons

In this section, we evaluate the efficiency of our score function with respect to the Wilcoxon scores for the generalized F distribution. In essence, we explore the selection of r and s that provides improvement over the Wilcoxon scores. For comparison purposes, ARE(11, rs) provided in Lemma 2 are calculated for right-skewed, left-skewed, light-tailed and heavy-tailed distributions respectively. We evaluated ARE(11, rs) for several values of $r, s = 0\,(3)\,0.1$. Figures depict a perspective plot of ARE(11, rs) as a function of r and s.

The generalized F distribution is a very flexible distribution that covers a variety of shape and tail behaviors. It produces symmetric distributions for $m_1 = m_2$, positively skewed distributions for $m_1 > m_2$, and negatively skewed distributions for $m_1 < m_2$. Further It produces heavy-tailed distributions for $m_1, m_2 < 1$, whereas it produces light-tailed distributions for $m_1, m_2 > 1$. McKean and Sievers(1989) adaptively estimated m_1 and m_2 to reflect on the shape of the underlying probability models. Namely the distribution of F is Weibull if $(m_1, m_2) \rightarrow (1, \infty)$, lognormal if $(m_1, m_2) \rightarrow (\infty, \infty)$, the generalized gamma if $(m_1, m_2) \rightarrow (\infty, 1)$.

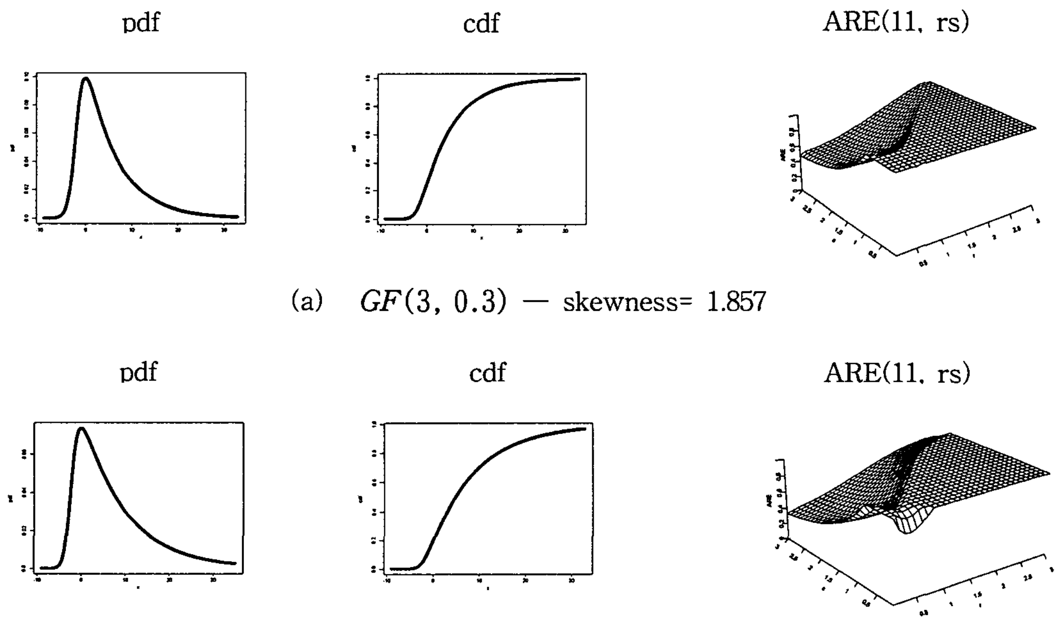


Plot 1. Schematic of the four classes, C1-C4, of the $GF(2m_1, 2m_2)$ distributions

In general this class of the generalized F distribution can be conveniently divided into the four subclasses C1 through C4 which are represented by the four quadrants with center $(m_1, m_2) = (1, 1)$ as depicted in Plot 1. When $m_1 = m_2 = 1$, the generalized F distribution, the point $(1, 1)$ in this plot, corresponds to the logistic distribution, $GF(2,\ 2)$, and forms a natural center point for the diversified distributions. The distributions in C1 are suitable for positively skewed distributions with heavy right tails and moderate left tails, whereas the distributions in C3 are suitable for negatively skewed distributions with heavy left tails and moderate right tails. In addition the distributions in C2 are suitable for light-tailed distributions, whereas the distributions in C4 are suitable for heavy-tailed distributions.

4.1 Right-Skewed Generalized F Distribution

Figure 1 and Table 1 show the pdf and cdf and asymptotic relative efficiency for right-skewed distributions such as $GF(3,\ 0.3)$ and $GF(4,\ 0.2)$. The computations were made for all $r, s = 0\ (3)\ 0.1$, but we only report the selected values.



(a) $GF(3,\ 0.3)$ — skewness= 1.857

(b) $GF(4,\ 0.2)$ — skewness= 1.913

Figure 1. Class C1 distribution, the right-skewed distribution

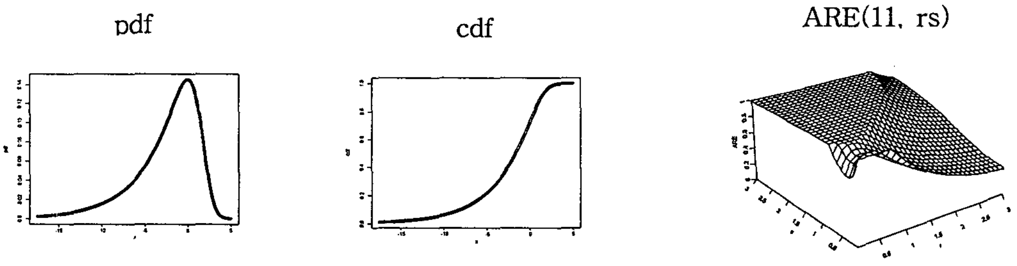
Table 1. ARE(11, rs) for the right-skewed distribution

r	0.1			0.5			0.9			1			2		3		skewness
distribution \ s	1	2	3	1	2	3	1	2	3	1	2	3	2	3	2	3	
GF(3, 0.3)	.934	.574	.477	.872	.614	.530	.961	.672	.576	1.000	.694	.593	1.000	.824	1.289	1.032	1.857
GF(4, 0.2)	.833	.460	.334	.771	.502	.391	.937	.595	.459	1.000	.627	.481	1.000	.716	1.250	.852	1.913

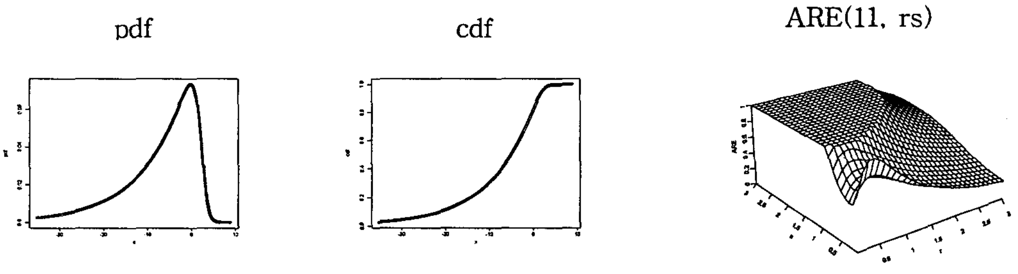
The results of Figure 1 and Table 1 can be summarized as follows. For the right-skewed distributions, they indicate that if $r < 1$ and $s > 1$, $v(\widehat{\beta_{r,s}})$ is much smaller than $v(\widehat{\beta_{1,1}})$. In particular, for a strongly right-skewed distribution, the proposed score generating function provides improved efficiency over the Wilcoxon score for the small values of r and large values of s. Thus for strongly right-skewed distributions with high positive skewness we select $r = 0.1$ and as large an s as possible.

4.2 Left-Skewed Generalized F Distribution

Figure 2 and Table 2 show the pdf and cdf and asymptotic relative efficiency for left-skewed distributions such as $GF(0.5, 6)$ and $GF(0.2, 4)$.



(a) $GF(0.5, 6)$ — skewness= -1.751



(b) $GF(0.2, 4)$ — skewness= -1.913

Figure 2. Class C3 distribution, the left-skewed distribution

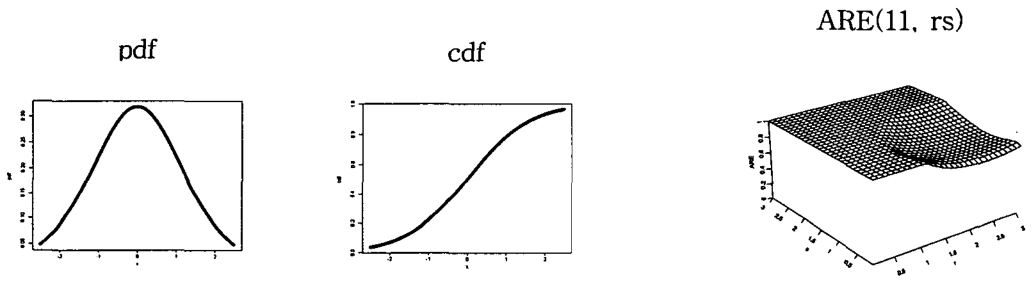
Table 2. ARE(11, rs) for the left-skewed distribution

distribution \ s	0.1			0.5			0.9			1			2		3		skewness
r	1	2	3	1	2	3	1	2	3	1	2	3	2	3	2	3	
GF(0.5, 6)	.808	.448	.325	.754	.493	.383	.934	.593	.456	1.000	.625	.478	1.000	.712	1.244	.843	-1.751
GF(0.2. 4)	.674	.348	.225	.634	.392	.278	.903	.524	.362	1.000	.566	.387	1.000	.611	1.173	.675	-1.913

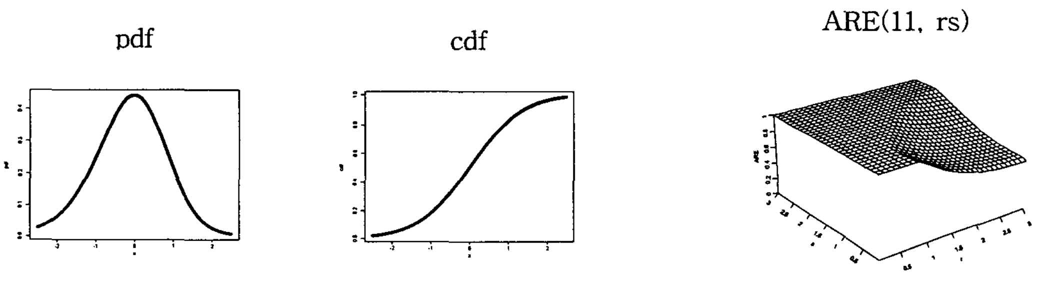
As we compare the left-skewed distributions with the right-skewed distributions, similar results can be observed. Figure 2 and Table 2 show that our procedure has higher efficiency than the Wilcoxon scores if $r > 1$ and $s < 1$. We should choose as large an r as possible and as small an s as possible. Thus, for strongly left-skewed distributions with low negative skewness we select $s = 0.1$ and as large an r as possible.

4.3 Light-Tailed Generalized F Distribution

Figure 3 and Table 3 show the pdf and cdf and asymptotic relative efficiency for light-tailed distributions such as $GF(3, 3)$ and $GF(4, 8)$.



(a) $GF(3, 3)$ — kurtosis= 3.806



(b) $GF(4, 8)$ — kurtosis= 3.625

Figure 3. Class C2 distribution, the light-tailed distribution

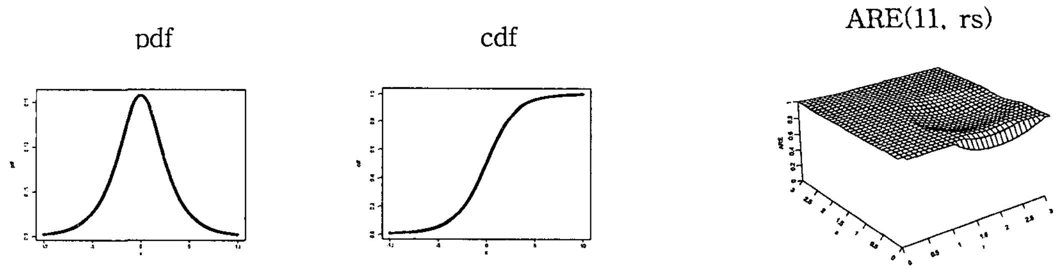
Table 3. ARE(11, rs) for the light-tailed distribution

distribution \ r / s	1.5			2			2.5			3			kurtosis
	0	1	2	0	1	2	0	1	2	0	1	2	
Logistic	.816	.895	1.063	.765	.833	1.000	.784	.856	.994	.854	.885	1.025	
GF(3, 3)	.798	.884	1.080	.730	.833	1.000	.729	.823	.979	.772	.839	.996	3.806
GF(4, 8)	.733	.844	1.145	.616	.758	1.000	.565	.714	.928	.550	.697	.900	3.625

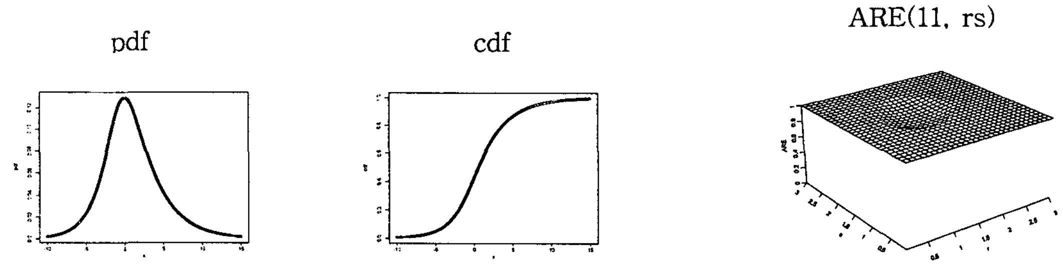
Figure 3 and Table 3 show that our estimator performs better than Wilcoxon score rank estimator for $1 < r < 3$ and $0 < s < 2$. Especially our procedure has best efficiency for $r = 3$ and $s = 0$ with low kurtosis. Some of the selected values of ARE(11, rs) are given in Table 3 because of the space limitation.

4.4 Heavy-Tailed Generalized F Distribution

Figure 4 and Table 4 show the pdf and cdf and asymptotic relative efficiency for heavy-tailed distributions such as $GF(1,\ 1)$ and $GF(1,\ 0.6)$.



(a) $GF(1,\ 1)$ — kurtosis= 5.000



(b) $GF(1,\ 0.6)$ — kurtosis= 5.848

Figure 4. Class C4 distribution, the heavy-tailed distribution

Table 4. ARE(11, rs) for the heavy-tailed distribution

distribution \ r, s	1.1	1.3	1.5	1.7	1.9	kurtosis
Cauchy	.989	.978	.976	.981	.992	
GF(1, 1)	.978	.953	.948	.958	.982	5.000
GF(1, 0.6)	.976	.949	.943	.954	.981	5.848

From Figure 4 and Table 4, we observe that for heavy-tailed distributions, values of $1 < r, s < 2$ yield higher asymptotic relative efficiencies than the Wilcoxon scores. In particular, when r and s are 1.5, $v(\widehat{\beta_{r,s}})$ is the minimum among all $r, s = 0\,(3)\,0.1$ that we computed. Thus for long heavy-tailed distributions with high kurtosis, we select $r, s = 1.5$.

5. Selection of r and s

We preliminarily suggest using the Wilcoxon scores to get residuals to calculate the size of skewness of the underlying probability model. Next in order to find a defining association between the degree of skewness, denoted by *Skew*, and the selection of m_1 and m_2 for the generalized F distribution, the following prediction fits are suggested. For the positive skewness,

$$m_1 = 1.2\,Skew - 0.3, \qquad m_2 = 0.5 - 0.2\,Skew$$

and for the negative skewness,

$$m_1 = 0.5 - 0.2\,|Skew|, \qquad m_2 = 1.2\,|Skew| - 0.3.$$

Further we finally fit a regression model to predict the optimal values of r and s for a given values of skewness. For the positive skewness,

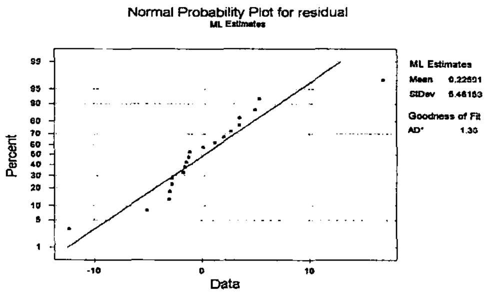
$$r = 1 - 0.5\,Skew, \qquad s = 1 + 0.95\,Skew$$

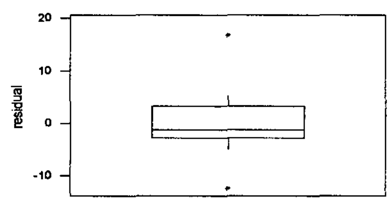
and for the negative skewness,

$$r = 1 + 0.95\,|Skew|, \qquad s = 1 - 0.5\,|Skew|.$$

Numerical Example. Consider the multiple regression model in Example 5.6 of Bowerman et al.(1986, p162) for sales associated with income and space. Residuals using the Wilcoxon scores are available in the command RREGRESS in Minitab with normal probability plot and box plot as follows.

1.2509	0.1744	4.9970	3.5353	2.1116
-2.7949	16.8501	-2.9724	-2.7951	-1.7339
-1.1345	5.3730	-3.0123	3.5412	-1.5658
-1.2509	-1.4191	-5.0735	-12.3120	2.7491





(1) Then we obtain $Skew = 0.888$ by using the UNIVARIATE procedure in SAS for the Wilcoxon residuals, for which we reject the null hypothesis in favor of the weakly right-skewed alternatives. (2) In addition from the formula provided, we find that $m_1 = 0.8$ and $m_2 = 0.3$. It makes us estimate the type of the underlying distribution, $GF(2m_1, 2m_2) = GF(1.6, 0.6)$. (3) We also predict the values of $r = 0.6$ and $s = 1.8$ even for a given value of mild skewness. (4) Finally the asymptotic relative efficiency of the proposed scores with respect to the Wilcoxon scores indicates that $ARE(11, rs) = 0.935$ for $r = 0.6$, $s = 1.8$ and $GF(1.6, 0.6)$ correspondingly.

6. Conclusions

In this paper we derived the asymptotic relative efficiency, ARE(11, rs), of our score function with respect to the Wilcoxon scores for the generalized F distributions. We thoroughly explored the selection of r and s of our new score function that provides improvement over the Wilcoxon scores. The result for asymmetric distributions which we encounter in practice commonly can be summarized as follows. We select $r < 1$, $s > 1$ for right skewed distribution and $r > 1$, $s < 1$ for left skewed distributions. In addition, we select $r = 3$, $s = 0$ for light tailed distribution and $r, s = 1.5$ for heavy tailed distributions.

References

[1] Ahmad, I. A. (1996). A Class of Mann-Whitney-Wilcoxon Type Statistics, *The American Statistician*, 50, 324-327.

[2] Bowerman, B. L., O'Connell, R. T. and Dickey, D. A. (1986). *Linear Statistical Models,* Duxbury Press, Boston.

[3] Choi, Y. H. and Ozturk, O (2002). A New Class of Score Generating Functions for Regression Models, *Statistics & Probability Letters*, 57, 205-214.

[4] Choi, Y. H. (2004). Asymptotic Relative Efficiency for New Score Functions in Rank Regression Models, *The Korean Journal of Applied Statistics,* 17, 269-280.

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

[6] Jaeckel, L. A. (1972). Estimating Regression Coefficients by Minimizing the Dispersion of Residuals, *The Annals of Mathematical Statistics*, 43, 1449-1458

[7] McKean, J. W. and Sievers, G. L. (1989). Rank Scores Suitable for Analyses of Linear Models under Asymmetric Error Distributions, *Technometrics*, 31, 207-218.

[8] Ozturk, O (1999). Two-Sample Inference Based on One-Sample Ranked Set Sample Sign Statistics, *Journal of Nonparametric Statistics*, 10, 197-212.

[9] Ozturk, O (2001). A Generalization of Ahmad's Class of Mann-Whitney-Wilcoxon Statistics, *Australian and New Zealand Journal of Statistics*, 43, 67-74.

[10] 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.

[11] Ozturk, O and Hettmansperger, T. P. (1997). Generalized Weighted Cramer-Von Mises Distance Estimators, *Biometrika*, 84, 283-294.

[Received January 2004 , Accepted August 2004]