J. King Saud Univ., Vol. 19, *Admin. Sci.* (1), pp. 41-50, Riyadh (1427H./2006)

A Robust Measure for the Correlation Coefficient

Moawad El-Fallah Abd El-Salam
Department of Psychology, College of Social Sciences,
Al-Imam Muhammad Ibn Saud Islamic University,
Riyadh, Saudi Arabia

(Received 5/2/1426H.; accepted for publication 1/1/1427H.)

Abstract. In this paper, we investigate the robustness of some well known correlation coefficients, namely, Pearson's, Spearman's and Kendall's. The empirical evidence shows that these correlation coefficients are sufficiently non-robust against outliers. That is, they do not have high breakdown points. As an alternative, a robust estimator for the correlation coefficient is proposed. This estimator is based on the least median of squares. It is shown that this correlation coefficient has a higher breakdown point than the well known correlation coefficients.

Keywords: Correlation coefficient, Outliers, Robustness, Least median of squares, High breakdown point.

1. Introduction

The correlation coefficient is a standard tool in applied regression analysis. Although it is not always thoughtfully used, it remains an informative summary measure of the predictive power of the selected regression model. However, since the least squares regression analysis is very sensitive to outliers, it is not surprising that the coefficient of correlation inherits this problem.

Let (x_1, y_1), ……., (x_n, y_n) be n observations from a bivariate normal distribution with parameters $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$, where μ_x and σ_x^2 are the mean and variance of x, μ_y and σ_y^2 are the mean and variance of y. ρ is the correlation coefficient between x and y which is given by $\rho=\beta \sigma_x / \sigma_y$, where β is the slope parameter of regression y on x. The sample correlation coefficient commonly used for estimating ρ is the Pearson's correlation coefficient which is defined as:

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{[\sum (y_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2]^{\frac{1}{2}}} \quad (1)$$

In regression analysis, it is often recognized that outliers can occur in either the dependent variable y or independent variable x or in both variables. The correlation coefficient r_p in (1) is based on the sample means $\bar{x}$ and $\bar{y}$, respectively, which are known to be very sensitive to the presence of outliers. The statistic which measures the effect of a possible outlier (x, y) at the correlation coefficient is called the influence function. In this respect, Romanazzi [1] illustrated the non-robustness of r_p by showing that its influence function is unbounded.

As an alternative to the Pearson's correlation coefficient, r_p , we may turn to non-parametric correlation coefficients which are based on the ranks of the observations. Two well known correlation coefficients of this type are Spearman's rho and Kendall's tau. Spearman's rho, r_s, can be computed as:

$$r_s = 1 - \frac{6D^2}{n(n^2 - 1)} \quad (2)$$

,

where,

$$D^2 = \sum_i^n (r_{y_i} - r_{x_i})^2 \quad ,$$

and r_{y_i} , r_{x_i} are the ranks of y_i and x_i respectively. Kendall's tau, r_k, is given by:

$$r_k = 1 - \frac{4Q}{n(n-1)} \quad , \quad (3)$$

where Q is the number of inversions between the rankings of x and y. An inversion is any pair of objects (i, j) such that $r_i - r_j$ and $r_i^{/} - r_j^{/}$ have opposite signs.

By replacing the values of the observations by their ranks, the effect of some extreme observations may be reduced. Therefore, we can expect that the rank correlation coefficients, r_s and r_k would be less sensitive to the outliers than the Pearson's correlation coefficient r_p. However, the question still arises here is which one of these rank correlations is sufficiently robust to a substantial amount of outliers, i.e. which one has high breakdown point. The breakdown point, BP, is the maximal fraction of outliers that an estimator can withstand or the smallest proportion of bad observations that the

estimator can resist before it breaks down. BP = 50% is the best that can be expected to be achieved by an estimator, (Rousseeuw and Leroy [2, p. 9]).

In this paper, we introduce a robust estimator of the correlation coefficient which is expected to have a high breakdown point. It is based on the least median of squares regression procedure. This robust correlation coefficient will be introduced in Section 2. Section 3 presents an illustrative example to illustrate the robustness of the correlation coefficient and its comparison with the non-robust correlation coefficients, r_p, r_s and r_k. Section 4 describes a simulation study in which the four correlation coefficients are compared in terms of their empirical breakdown points, and other sampling properties, namely the bias, variance, standard error, mean square error and root mean square error will also be presented. Section 5 contains the conclusion.

2. A Robust Correlation Coefficient

Consider the following linear regression model:

$$y_i = \beta_o + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \in_i \ , \qquad i = 1,2, \ldots, n \tag{4}$$

An estimator of β_j ($j = 0, \ldots, k$) with 50% BP is the Least Median of Squares (LMS). It is defined as the value which minimizes:

$$\text{med } e_i^2 \quad , \tag{5}$$

where:

$$e_i = y_i - \sum_j^k \hat{\beta}_j x_{ij} \ , \qquad i = 0, 1, \ldots, n$$

The LMS performs poorly when the errors are really normally distributed. To overcome this problem, one should combine the LMS estimator with an efficient maximum likelihood type estimator (M-estimator) in the following way. Start with the LMS estimator and iterate with a redescending M-estimator (Hampel *et al.*, [3]).

Alternatively, Rousseeuw and Leroy [2, p. 76] suggested that, one can apply a weighted least squares defined by:

$$\underset{\hat{\beta}}{\text{Minimize}} \sum_i^n w_i e_i^2 \quad , \tag{6}$$

where:

$$w_i = \begin{cases} 1 & \text{if} \quad \left|\frac{e_i}{s}\right| \le 2.5 \\ 0 & \text{if} \quad \left|\frac{e_i}{s}\right| > 2.5 \end{cases}, \tag{7}$$

$$s = 1.4826 \left[1 + \frac{5}{n-k}\right] \sqrt{\text{med}\, e_i^2} \tag{8}$$

This means simply that observation i will be retained in the weighted least squares if its absolute standardized residual is reasonably small or moderate, but omitted if it is an outlying observation. The criterion (7) may be interpreted as a 'hard' rejection of outliers in which only 'good' observations are retained in the data set. Therefore, the resulting estimator will still possess the high breakdown point, but is more efficient under the normality assumption.

The robust correlation coefficient between x and y associated with (6) and (7) is now defined as:

$$r_w = \frac{\sum w_i (x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\left[\sum w_i (\bar{x}_i - \bar{x}_w)^2 \sum w_i (\bar{y}_i - \bar{y}_w)^2\right]^{\frac{1}{2}}}, \tag{9}$$

where :

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad , \quad \text{and} \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i},$$

With w_i being defined by (7), r_w can then be expected to have the same maximal breakdown point, i.e. BP = 50% as possessed by the LMS regression estimator. Also, r_w can be viewed as the weighted Pearson's correlation coefficient which considers only the 'good' observations in the data set.

3. Illustrative Example

In order to compare the various correlation coefficients, we consider the so-called Pilot-plant chemical data from Daniel and Wood [4, p. 46]. The response variable (y) corresponds to acid content determined by titration and the explanatory variable (x) is the organic acid content determined by extraction and weighting (the data is presented in the appendix).

Considering the data assuming that one of the observations, i.e. the x-value of the sixth observation has been wrongly recorded as 370 instead of 37. The results of various

correlation coefficients based on the data which consists of n = 20 observations are presented in Table 1.

Table 1. The values of r_p, r_s. r_k and r_w for the Pilot-plant data

Correlation coefficient		$x_6 = 370$	$x_6 = 37$
Pearson's	r_p	0.38	0.99
Spearman's	r_s	0.76	0.99
Kendall's	r_k	0.73	0.90
LMS	r_w	0.99	0.99

The results show that Pearson's correlation coefficient has been strongly affected by the single outlier. The Spearman's and Kendall's correlation coefficients seem to be slightly affected by the extreme observation. On the other hand, the LMS based correlation coefficient was not at all influenced by the outlier because it was based on the 'reduced or clean' observations with nonzero weights.

4. Simulation Study

We carry out a simulation study to illustrate the breakdown properties of the correlation coefficients; r_p , r_s , r_k and r_w as were defined by (1), (2), (3) and (9). We begin with generating 100 'good' observations according to the linear relation: $y_i = 2.0 + 1.0\ x_i + \varepsilon_i$, where x_i is normally distributed with mean 5 and variance 1, ε_i is drawn from $N(0, \sigma^2)$, $\sigma = 0.2$ and the true value of ρ is 1.0. The normal varieties were generated by the NAG program subroutine G05DDF on the IBM 4341 computer system. Using these data, we applied Eqs. (1), (2), (3) and (9). The obtained results are: $r_p = 0.984$, $r_s = 0.976$, $r_k = 0.876$ and $r_w = 0.987$. Because the data were uncontaminated, all the correlation coefficients yielded values which are close to the original $\rho = 1$.

Then, we start to contaminate the data. At each step, we delete one 'good' observation and replace it with a 'bad' data point. The contaminated data point was generated according to the linear relation where x_i is uniformly distributed on (5, 10) and y_i is normally distributed with mean 2 and standard deviation 0.2. This is repeated until only 50 'good' observations remained. Table 2 presents the values of r_p, r_s, r_k and r_w when 'good' observations are replaced by a certain percentage of outliers.

Table 2. The values of r_p, r_s, r_k and r_w for n = 100 and $\rho = 1$

Contamination (%)	r_p	r_s	r_k	r_w
0	0.984	0.976	0.876	0.987
10	- 0.070	0.503	0.547	0.987
20	- 0.317	0.195	0.314	0.988
30	- 0.451	-0.091	0.096	0.988
40	- 0.603	- 0.380	- 0.119	0.983
45	- 0.605	- 0.448	- 0.117	0.985
50	- 0.601	- 0.489	- 0.225	- 0.710

From the results of Table 2, we see that r_p was immediately affected by outliers and its value moves away from the true value as the percentage of outliers increases. It can be noted that r_p breaks down first and is then followed by the rank correlation coefficients r_s and r_k. The increase in the percentage of outliers from 0% (no contamination) up to 45% contamination has changed not only the values but also the signs of r_p, r_s and r_k, i.e. from the positive to the negative values of the correlation coefficients. It appears that the LMS-based correlation coefficient r_w holds on before breaking down at 50% of the outliers.

The breakdown properties of these sample correlation coefficients are investigated further by looking at five summary statistics, namely, the bias, variance, standard error (SE), mean square error (MSE), and root mean square error (RMSE) in 500 trials. In each trial t (t = 1, 2, …, 500), a sample of size 20, 50 and 100, respectively, was generated according to the sampling situations described earlier. The average of the sample correlation coefficient $\hat{\rho}$ is $\bar{\rho} = T^{-1} \Sigma \hat{\rho}_t$ which yields the bias $\bar{\rho} - \rho$. The variance is given by $v(\hat{\rho}) = T^{-1} \Sigma(\hat{\rho}_t - \bar{\rho})^2$ which can be used to compute the MSE as: $MSE(\hat{\rho}) = [bias]^2 + v(\hat{\rho})$. Accordingly, the SE is given by $\sqrt{v(\hat{\rho})}$ and the RMSE by $\sqrt{MSE(\hat{\rho})}$. These summary statistics are presented in Table 3 for n = 20, 50 and 100.

Table 3. Summary statistics for r_p, r_s, r_k and r_w for n = 20, 50, 100 and $\rho = 1$

Contamination	Correlation	n = 20			n = 50			n = 100		
(%)	coefficient	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
0	r_p	–0.21	0.012	0.024	–0.020	0.006	0.020	–0.020	0.004	0.020
	r_s	–0.034	0.021	0.041	–0.027	0.010	0.029	–0.025	0.006	0.025
	r_k	–0.121	0.046	0.130	–0.125	0.023	0.127	–0.126	0.015	0.127
	r_w	–0.017	0.012	0.021	–0.018	0.006	0.020	–0.017	0.004	0.018
10	r_p	–1.112	0.254	1.140	–1.097	0.156	1.108	–1.102	0.102	1.107
	r_s	–0.501	0.087	0.508	–0.491	0.056	0.494	–0.490	0.036	0.491
	r_k	–0.453	0.056	0.456	–0.447	0.033	0.448	–0.446	0.021	0.446
	r_w	–0.018	0.012	0.021	–0.018	0.006	0.019	–0.018	0.004	0.018
20	r_p	–1.380	0.187	1.392	–1.371	0.120	1.376	–1.374	0.081	1.376
	r_s	–0.873	0.101	0.879	–0.864	0.068	0.867	–0.865	0.045	0.866
	r_k	–0.732	0.062	0.735	–0.718	0.039	0.719	–0.715	0.025	0.715
	r_w	–0.019	0.012	0.022	–0.018	0.007	0.019	–0.018	0.004	0.019
30	r_p	–1.503	0.152	1.511	–1.493	0.092	1.496	–1.492	0.068	1.493
	r_s	–1.159	0.119	1.165	–1.149	0.073	1.151	–1.150	0.054	1.151
	r_k	–0.954	0.076	0.957	–0.935	0.043	0.936	–0.931	0.031	0.932
	r_w	–0.020	0.013	0.024	–0.019	0.008	0.020	–0.018	0.005	0.019
40	r_p	–1.510	0.123	1.575	–1.546	0.080	1.548	–1.550	0.059	1.551
	r_s	–1.362	0.117	1.367	–1.347	0.075	1.349	–1.347	0.055	1.348
	r_k	–1.123	0.079	1.126	–1.100	0.047	1.101	–0.095	0.033	1.096
	r_w	–0.021	0.014	0.026	–0.019	0.008	0.020	–0.019	0.005	0.019
50	r_p	–1.582	0.169	1.587	–1.580	0.077	1.582	–1.574	0.050	1.574
	r_s	–1.473	0.123	1.478	–1.474	0.078	1.476	–1.465	0.054	1.466
	r_k	–1.235	0.087	1.238	–1.220	0.052	1.221	–1.206	0.036	1.206
	r_w	–1.664	0.136	1.670	–1.659	0.084	1.661	–1.650	0.098	1.653

From the results of Table 3, we see that the r_p and r_w correlation coefficients provide the best results when no contamination occurs in the model. As a result, the rank correlation coefficients r_s and r_k perform somewhat less than the r_p and r_w based counterparts in the normal situation. It can be noted that the bias is negligible in this situation and the variance makes up most of the MSE. In addition, as the percentage of outliers increases in the data, the r_p, r_s and r_k correlation coefficients break down systematically at these contaminated samples and they have very high MSE values. For these correlation coefficients, the bias makes up most of the MSE.

On the other hand, the LMS-based correlation coefficient performs reasonably well even in the situation where there are nearly 50% outliers in the data. The results seem to be consistent in all 500 trials and for each sample size n = 20, 50 and 100. Therefore, we conclude that the LMS-based correlation coefficient, r_w, has a higher breakdown point than the usual correlation coefficients, r_p, r_s and $r_{k,}$ because it is able to withstand substantial amounts of outliers in the data.

5. Conclusion

We have seen that the Pearson's correlation coefficient is very sensitive to the presence of outliers. The empirical study shows that the Spearman's and Kendall's rank correlation coefficients are not sufficiently robust when the percentage of outliers increases in the data set. Therefore, they cannot provide a robust alternative to the Pearson's correlation coefficient. In this respect, the LMS-based correlation coefficient indeed achieves the goal for which it was constructed because it is able to produce satisfactory results even in the presence of a large amount of outliers.

Appendix

The Pilot–plant data

Obs.	y	x
1	76	123
2	70	109
3	55	62
4	71	104
5	55	57
6	48	37
7	50	44
8	66	100
9	41	16
10	43	28
11	82	138
12	68	105
13	88	159
14	58	75
15	64	88
16	88	164
17	89	169
18	88	167
19	84	149
20	88	167

Source: Daniel and Wood [4, p. 46].

References

[1] Romanazzi, M. “Influence in Canonical Correlation Analysis.” *Biometrika*, 57 (1992), 237-259.
[2] Rousseeuw, P.J. and Leroy, A.M. *Robust Regression and Outlier Detection.* New York: John Wiley, 1987.
[3] Hampel, F.R.; Rousseeuw, P.J. and Ronchetti, E. “The Change of Variance Curve and Optimal Redescending M-estimators.” *J.A.S.A.*, 76 (1981), 643-648.
[4] Daniel, C. and Wood, F.S. *Fitting Equations to Data.* New York: John Wiley, 1980.

مقياس بديل لمعامل الارتباط

معوض بن الفلاح عبد السلام

أستاذ مساعد إحصاء، قسم علم النفس، كلية العلوم الاجتماعية،
جامعة الإمام محمد بن سعود الإسلامية، الرياض، المملكة العربية السعودية

(قدم للنشر في ١٤٢٦/٢/٥هـ؛ وقبل للنشر في ١٤٢٧/١/١هـ)

ملخص البحث. إن وجود قيم متطرفة في البيانات يؤثر على تقدير معالم نموذج الانحدار والإحصاءات المرتبطة بها، ومنها قيمة معامل الارتباط، ومن ثم الوصول إلى نتائج غير دقيقة، لذا يقترح استخدام بعض الطرق المقاومة (Robust) لتقليل أثر وجود القيم المتطرفة في البيانات.

وتهتم هذه الدراسة بإيجاد تقديرا بديلا لمعاملات الارتباط البسيطة المعروفة وذلك حالة احتواء البيانات على قيم متطرفة. ويعتمد المقياس البديل المقترح على إحدى الطرق المقاومة لآثار القيم المتطرفة. ولقد تبين من نتائج استخدام بعض البيانات الحقيقية وبيانات المحاكاة أن المقياس البديل المقترح للارتباط يتميز بخصائص إحصائية أفضل - في حالة وجود قيم متطرفة - من المعاملات المعروفة للارتباط.