ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(1.11)

Measuring the Relationship of Bivariate Data Using Hodges-Lehman Estimator

Suhaida Abdullah*, Nur Amira Zakaria, Nor Aishah Ahad, Norhayati Yusof and Sharipah Soaad Syed Yahaya

College of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Malaysia

The relationship of bivariate data ordinarily measured using correlation coefficient. The most commonly used correlation coefficient is the Pearson correlation coefficient. This coefficient is well-known as the best coefficient for interval or ratio bivariate data with a linear relationship. Even though this coefficient is good under the mentioned condition, it also becomes very sensitive to a small departure from linearity. Usually, this is because of the existence of an outlier. For that reason, this paper provides new robust correlation coefficients which combine the elements of nonparametric technique from the Hodges Lehmann estimator and the parametric technique based on the Pearson correlation coefficient. This paper also introduces different scale estimators such as median and median absolute deviation (MAD_n) and denoted by $r_{HL(med)}$ and $r_{HL(MADn)}$ respectively. The performance of the proposed correlation coefficients is measured by the coefficient values and these values are also being compared to the Pearson correlation coefficient and several existing robust correlation coefficients. The results show that the Pearson correlation coefficient (r) with no doubt is very good under perfect data condition, but with only 10% outliers, it not only give poor correlation value but turns the direction of the relationship to negative. While the $r_{HL(med)}$ and $r_{HL(MADn)}$ offer the highest coefficient values and these values are robust to the existence of outliers by up to 30%. With very good performance under all data conditions yet simple in the calculation, the $r_{HL(med)}$ and $r_{HL(MADn)}$ is considered a good alternative to the r when need to deal with outliers.

Keywords: correlation coefficient; Hodges Lehmann; median; median absolute deviation (MADn)

I. INTRODUCTION

The correlation coefficient is a known coefficient to measure a relationship between two variables. Pearson correlation coefficient is one of the most commonly used correlation coefficients especially when the variables having a linear relationship, but it becomes poor when the relationship deviates from linearity. This shortcoming is usually handled by using nonparametric correlation coefficients such as Spearmen or Kendal Tau correlation coefficient. These correlation coefficients have not influenced by the presence of the outlier due to the uses of rank in their calculation. However, rank is not the best option to avoid the effect of the outlier because it does not use the original data. As stated by Xu *et al.,* (2016) using rank instead of the original data might lead to the losing of useful information.

The Pearson correlation coefficient unable to handle the outlier due to the use of the mean as its location estimator. Mean is known to be very sensitive to the outlier with 0% breakdown point. This drawback encourages the development of a robust correlation coefficient as alternatives to the Pearson correlation coefficient in handling the outlier. The robust correlation coefficient can be a better option compared to the nonparametric because it lessens the influence of the outlier but remains to use the original data.

To date, the robust correlation coefficient base on median developed by Sheylyakov *et. al.,* (2012) provided a more reliable measurement of the coefficient. Median is known to have the maximum breakdown point which is 50%. However,

*Corresponding author's e-mail: suhaida@uum.edu.my

the more robust estimator not always be the best estimator. The efficiency of the estimator also plays an important role in order to provide better properties to the coefficient. The more robust the estimator will reduce the efficiency of the estimator (Geyer, 2003).

Hence, in choosing a suitable estimator for developing any coefficient measure, the efficiency also needs to be considered. Besides the mean and median, Hodges Lehman (*HL*) estimator is a worth estimator to study on. The investigation on the efficiency of the *HL* estimator revealed that this estimator is more efficient compared to mean and median under most conditions of *t*-distribution family. It also has an intermediate breakdown point with 30%.

Based on the good properties of the *HL* estimator, therefore, the objective of this paper is to develop a robust correlation coefficient using the *HL* estimator which believe will improve the performance of the correlation coefficient in measuring the relationship of two variables. The evaluation of the developed robust correlation coefficient is assessed based on the simulation study and to check the validity, real data analysis is conducted.

II. MATERIAL AND METHODS

The development of the robust correlation coefficient using the *HL* estimator in this study is based on work by Sheylyakov *et. al.,* (2012). Their robust correlation coefficient utilizes median absolute deviation (*MAD*) as location and scale estimator to obtain a median correlation coefficient and *MAD* correlation coefficient as defined in equation 1 and 2.

$$r_{med} = \frac{\left(med^2|u| - med^2|v|\right)}{\left(med^2|u| + med^2|v|\right)} \quad (1)$$

$$r_{MAD} = \frac{(MAD^2(u) - MAD^2(v))}{(MAD^2(u) + MAD^2(v))} \quad (2)$$

where

$$u = \frac{x - med(x)}{\sqrt{2}MAD(x)} + \frac{y - med(y)}{\sqrt{2}MAD(y)} \quad (3)$$

$$v = \frac{x - med(x)}{\sqrt{2}MAD(x)} - \frac{y - med(y)}{\sqrt{2}MAD(y)} \quad (4)$$

The calculation of this coefficient is based on a robust scale estimator namely median absolute deviation (*MAD*). The formula for *MAD* estimator is shown in equation 5.

$$MAD = med|X_i - medX| \quad (5)$$

The *MAD* was promoted by Hampel (1974) with maximum breakdown point which is 50% and bounded influence function. These properties increase the ability of the correlation coefficient in handling outlier. Based on work by Sheylyakov *et. al.,* (2012), the *MAD* provides more robust result under contaminated data especially when the sample size is small. They also found that the *MAD* can be an efficient scale estimator and suitable to be used in measuring dispersion (in equation 3 and 4). Thus, in the development of a robust correlation coefficient using the *HL* estimator, the *MAD* is remained as scale estimator as in equation 3 and 4.

The *HL* estimator was first introduced by Hodges and Lehmann (1963) where it found to be a consistent and median-unbiased estimator of the population mean under symmetric distribution. This estimator also estimates the "pseudo-median" that is closely related to population median (Boos, 1982) under non-normal distribution. Equation 6 describes the calculation of the *HL* estimator.

$$\hat{\theta} = median\left\{\frac{X_i + X_j}{2}; 1 \le i \le j \le n\right\} \quad (6)$$

So, in this study, the robust correlation coefficient using the *HL* estimator is derived as:

$$r_{HL(med)} = \frac{\left(med|u|\right)^2 - \left(med|v|\right)^2}{\left(med|u|\right)^2 + \left(med|v|\right)^2} \quad (7)$$

with considering the median to measure the dispersion. While for *MAD* as the measurement of dispersion, the coefficient is denoted as:

$$r_{HL(MAD)} = \frac{\left(MAD(u)\right)^2 - \left(MAD(v)\right)^2}{\left(MAD(u)\right)^2 + \left(MAD(v)\right)^2} \quad (8)$$

For both equation (7) and (8) implied the same formula for u and v where the *HL* and *MAD* as its location and scale estimator respectively.

$$u = \frac{x - HL(x)}{\sqrt{2}MAD(x)} + \frac{y - HL(y)}{\sqrt{2}MAD(y)} \quad (9)$$

$$v = \frac{x - HL(x)}{\sqrt{2}MAD(x)} - \frac{y - HL(y)}{\sqrt{2}MAD(y)} \quad (10)$$

Besides the *MAD* as a scale estimator, this study also investigated the performance of the correlation coefficient when employed another robust scale estimator that is the *MADn*. The *MADn* is the *MAD* that multiply by a constant value b=1.4826 that made the *MAD* more consistent especially under asymmetric distribution. Therefore the $r_{HL(MADn)}$ is indicated as:

$$r_{HL(MADn)} = \frac{(MADn^2u - MADn^2v)}{(MADn^2u + MADn^2v)} \quad (11)$$

with u and v are denoted as

$$u = \frac{x - med(x)}{\sqrt{2}MAD_n(x)} + \frac{y - med(y)}{\sqrt{2}MAD_n(y)} \quad (12)$$

and

$$v = \frac{x - med(x)}{\sqrt{2}\ MAD_n(x)} - \frac{y - med(y)}{\sqrt{2}\ MAD_n(y)} \quad (13)$$

To evaluate the performance of the proposed robust correlation coefficients, a simulation data is used to get the coefficient values. This simulated data was set at prior to has perfect correlation where ρ = 1. Therefore, the nearest correlation value to 1 is considered the best. The perfect data condition is simulated using the linear equation of $y_i = 2.0 + 1.0x_i + u_i$ where $x_i \sim Normal(5,1)$ and $u_i \sim Normal(0, 0.04)$. To see the effect of the outlier, three percentages of contaminated data also included in the simulation study that are 10%, 30% and 50%. The contaminated data is performed by $y_i \sim Normal(2, 0.04)$ and $x_i \sim Uniform(5,10)$. For the sample sizes, this study investigates the performance of the proposed correlation coefficient values under small sample with n=25, a moderate sample with n=100 and large sample with n= 400.

III. RESULTS AND DISCUSSION

The performance of the proposed correlation coefficient which based on the coefficient values is as depicted in Table 1. The correlation coefficient values of the proposed methods also being compared with the other existing robust correlation coefficients such as the Pearson correlation coefficient (r), the correlation coefficient based on median which recommended by Sheylyakov *et. al.,* (2012) denoted by r_{med} and r_{MAD}. The proposed Hodges Lehmann correlation coefficients are denoted as $r_{HL(med)}$, $r_{HL(MAD)}$ and $r_{HL(MADn)}$ that employed the scale estimator median, *MAD* and *MADn* respectively.

Based on Table 1, under perfect data condition with 0% contamination, all correlation coefficient values perform well with the value that is almost 1. The Pearson correlation coefficient (r) is the best as it known. However, the r is very not robust where the values are all demolition when there is at least 10% contamination in the data. It does not only fail to measure the degree of relationship but mistakenly change the direction of the relationship to negative.

When there is a data contamination for at least 10%, the other robust correlation coefficients offer better measurement of relationship. Under 10% contamination, the $r_{HL(MADn)}$ has the best measurement with the nearest to 1 for small sample size. For a larger sample size, the r_{MAD} and $r_{HL(MAD)}$ perform best.

The r_{MAD} and $r_{HL(MAD)}$ also found to be the best correlation coefficient for a bigger percentage of contamination which up to 30%.

Table 1. The correlation coefficient values using simulated data

Data	Sample size (n)	r	r_{med}	r_{MAD}	$r_{HL(med)}$	$r_{HL(MAD)}$	$r_{HL(MADn)}$
Perfect Data	25	**0.9990**	0.9975	0.9984	0.9978	0.9984	0.9984
	100	0.9990	0.9991	**0.9993**	0.9992	**0.9993**	**0.9993**
	400	**0.9992**	0.9987	0.9987	0.9986	0.9987	0.9987
Contaminated Data (10%)	25	-0.1386	0.9723	0.9974	0.8715	0.9974	**0.9981**
	100	-0.2763	0.9368	**0.9972**	0.8724	**0.9972**	0.9962
	400	-0.1712	0.9318	**0.9966**	0.8621	**0.9966**	0.9941
Contaminated Data (30%)	25	-0.6885	0.0267	**0.8623**	-0.6611	**0.8623**	0.7309
	100	-0.5289	0.1214	**0.9097**	-0.4140	**0.9097**	0.8446
	400	-0.4846	0.2375	**0.9313**	-0.4200	**0.9313**	0.8345
Contaminated Data (50%)	25	-0.7320	-0.7278	0.3701	-0.7607	0.3701	-0.2373
	100	-0.6102	-0.6769	-0.5342	-0.7148	-0.5342	-0.5561
	400	-0.5892	-0.6444	-0.6078	-0.6070	-0.6078	-0.5776

Table 2. The number of people with no working experience and the number of unemployment based on states in Malaysia for the year 2014

	Number of people with no working experience	Number of unemployed	Number of unemployed (with added outlier)
Johor	**224**	**191**	**91**
Kedah	240	188	188
Kelantan	270	230	230
Melaka	91	68	68
Negeri Sembilan	126	100	100
Pahang	150	111	111
Perak	203	159	159
Perlis	38	34	34
Pulau Pinang	77	57	57
Selangor	227	164	164
Terengganu	139	115	115
- Kuala Lumpur	80	60	60
- Labuan	23	4	4
- Putrajaya	8	10	10

Source of data: Statistics Department of Malaysia

From all the values of the coefficients, it is noticeable that the value of the r_{MAD} and $r_{HL(MAD)}$ are all exactly the same. It is something interesting to study where the usage of different scale estimator might influence the robustness of the correlation coefficient. This can be seen in the change of the coefficient values of the proposed *HL* correlation coefficient based on the *MADn* as it scales estimator.

To validate the proposed *HL* correlation coefficient, this study also proceeds with the analysis of using real data. For this reason, a data set of the number of people with no working experience and the number of unemployment based on states in Malaysia for the year 2014 is used. The original data is as depicted in Table 2.

Based on Table 2, a scatter plot as depicted in Figure 1 reviews how the relationship between the number of people with no working experience and the number of unemployed based on states in Malaysia for the year 2014.

From Figure 1, it can be seen that most of the plots are on the straight line which can be considered as a strong linear relationship.

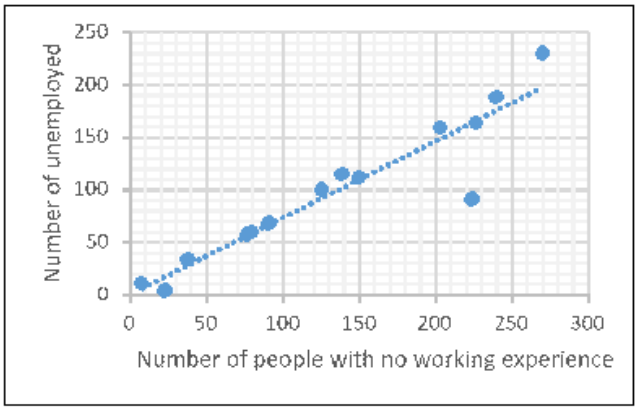

Figure 1

However, if the outlier exists, one value will deviate from the straight line. For example, data from Johor was modified to give the effect of outlier (from 191 to 91) as bolded in Table 2.

Table 3 displays the compared correlation coefficient values for this data. In Table 3, all coefficient values give highly correlation measurement with all are above 0.9. However, the $r_p$has the smallest value which projected that how it starts to be influenced by the outlier. Whereas, the other robust correlation coefficients have more than 0.95. the r_{med}, r_{MAD}, $r_{HL(MAD)}$ and $r_{HL(MADn)}$ have exactly the same value with 0.97833. This indicates that the proposed methods (the *HL* correlation coefficients) valid to be used for real data.

Table 3. Correlation coefficient values

Correlation coefficient	Coefficient value
r_p	0.93770
r_{med}	0.97833
r_{MAD}	0.97833
$r_{HL(MED)}$	0.95330
$r_{HL(MAD)}$	0.97833
$r_{HL(MADn)}$	0.97833

IV. CONCLUSION

In measuring the degree of relationship, the Pearson correlation coefficient is always the number one choice especially when the variables are known to have a linear relationship. When it comes to non-linear or if there is an outlier in the data set, the reliability of this coefficient totally

diminished by even a minimal number of the outlier.

The development of robust correlation coefficients offered a solution to this problem where the usage of the median in the correlation coefficient able to handle the occurrence to the outlier (Sheylyakov et al (2012)). With the highest breakdown point, the median is considered very robust but not always can be considered as the best estimator. When it takes into account the efficiency, the *HL* estimator seems to be more efficient (Geyer, 2003). Based on this point, the *HL* correlation coefficient provided another option to the Pearson correlation coefficient when it comes to the existence of the outlier. The study revealed that the performance of the $r_{HL(MAD)}$ exactly the same with $r_{(MAD)}$. But is some cases the $r_{HL(MADn)}$ provides a better result. It is interesting to further the investigation of the r_{HL} coefficient value using different scale estimator as promoted by Rousseeuw and Croux (1993).

V. ACKNOWLEDGEMENTS

We earnestly acknowledge the Universiti Utara Malaysia for the financial support under Universiti Grant Scheme (Code S/O 13377) and RIMC for facilitating the management of the research.

VI. REFERENCES

Abdullah, M. B. 1990. The Statistician 455-460.

Boos, D. D. 1982, A test for asymmetry associated with the Hodges-Lehmann estimator. Journal of the American Statistical Association, 77(379), 647-651.

Geyer, C. J. 2003, Nonparametric Tests and Confidence Intervals. Stat 5102 Notes. http://www.stat.umn.edu/geyer/old03/5102/notes/rank.pdf.

Hodges Jr, J. L. & Lehmann, E. L. 1963, Estimates of location based on rank tests. The Annals of Mathematical Statistics, 598-611 (1963).

Rousseeuw, P. J. and Croux, C. 1993, Journal of the America [n Statistical association 88(424), 1273-1283.

Shevlyakov, G. L. Smirnov, P. O. Shin, V. I. & Kim, K. 2012, Asymptotically minimax bias estimation of the correlation coefficient for bivariate independent component distributions. Journal of Multivariate Analysis, 2012, 111, 59-65.

Xu, W. Ma, R. Zhou, Peng Y. S. & Hou. Y. 2016, Asymptotic properties of Pearson′ s rank-variate correlation coefficient in bivariate normal model. Signal Processing, 119, 190-202.