A Robust Correlation Coefficient using LMS

Debarshi Chakraborty
M.Stat 2nd Year
Indian Statistical Institute, Kolkata

Roll No : MD2105
Course : Robust Statistics
Instructor : Dr. Ayanendranath Basu

April 14, 2023

Abstract

This review is based on the paper *On a Robust Correlation Coefficient* by Mokhtar Bin Abdullah published in the *Journal of the Royal Statistical Society. Series D (The Statistician), 1990.* A new correlation coefficient is proposed, which has a higher breakdown point compared to the well known correlation coefficients in literature. Comparative study is demonstrated via simulation.

Contents

1 Introduction

Suppose we have n realizations $(x_i, y_i)_{i=1}^n$ from (X, Y) having some joint distribution, where X and Y are continuous random variables. The population correlation coefficient between X and Y is given by

$$\rho = \frac{E(X-E(X))(Y-E(Y)))}{\sqrt{E(X-E(X))^2}\sqrt{E(Y-E(Y))^2}}$$

1.1 Pearson's product moment correlation coefficient

The sample counterpart of this quantity, known as the Pearson's product moment correlation coefficient is the most commonly used estimator of ρ. This is given by

$$r_P = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n (y_i-\bar{y})^2}}$$

Despite being a consistent estimator of ρ, r_P performs poorly in presence of outliers, which can occur in either the dependent variable y or the independent variable x. Intuitively, this is justified, since r_P is calculated using the sample means $\bar{x}$ and $\bar{y}$ respectively, which are known to be very sensitive in the presence of outliers and thus is highly non robust (low breakdown point, unbounded influence function). Moreover, it can be shown that the influence function of r_P is unbounded.

1.2 Alternative measures of correlation

To counter the problem of low breakdown, we may opt for nonparametric measures of correlation based on ranks, such as Spearman's rho (r_S) and Kendall's Tau (r_K).

Spearman's coefficient is given by

$$r_S = 1 - \frac{6d^2}{n(n^2-1)}$$

where $d^2 = \sum_{i=1}^n (r_{2i} - r_{1i})^2$ and r_{1i}, r_{2i} denote the ranks of x_i and y_i respectively.

Kendall's Tau is given by

$$r_K = 1 - \frac{4q}{n(n-1)}$$

where q is the number of inversions between the rankings of x data and y data.

Using the ranks of the observations instead of their exact values reduce the effect of extreme values to some extent. Thus, it is expected that r_S and r_K will be less sensitive to outliers compared to r_P.

1.3 Breakdown of an estimator of correlation

The question is, whether they have high breakdown points i.e. whether the coefficients are robust enough to a substantial amount of outliers. According to Donoho and Huber, any estimator of correlation coefficient which lies in the interval $[-1, 1]$ breaks down when the estimate can be moved to either endpoint of the interval or a nonzero correlation can be driven to zero by the contamination. A breakdown point of maximum ϵ^* =50% can be expected, because if we contaminate more than half of the data, the "good" and "bad" parts become indistinguishable.

2 The correlation coefficient

In the paper, the author proposes a correlation coefficient with high breakdown point, motivated by the least median of squares (LMS) and weighted least squares regression procedure.

2.1 Idea of LMS and WLS

For a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \;\; \forall i = 1, 2, ..., n$$

a robust estimator is the LMS estimator (Rousseeuw [Rou84]) defined as the value of $(\beta_0, \beta_1)^T$ which minimizes median(e_i^2) where $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \;\; \forall i = 1, 2, ..., n$.

Although is has a breakdown point of $\epsilon* = 50\%$, the LMS performs poorly when the errors are actually normally distributed. To get rid of this problem, one approach can be combining the LMS estimator with an efficient M-estimator. An alternative way suggested by Rousseeuw and Leroy (1987) is to apply a weighted least squares defined by

$$(\hat{\beta}_0, \hat{\beta}_1)\text{=argmin} \textstyle\sum_{i=1}^{n} w_i e_i^2$$

where $w_i = I(|e_i/\hat{\sigma}| \leqslant 2.5)$. In a nutshell, the i^{th} observation is retained only if its absolute standardized residual is not large i.e. reasonably small or moderate. Since, we are considering only the "good" points present in the data, the resulting estimator still continues to have a high breakdown point, moreover is more efficient than LMS under normality of errors.

2.2 The robust estimator of correlation

The author defines (Abdullah [Abd90]) a robust coefficient of correlation between x and y as

$$r_w = \frac{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w)}{\sqrt{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2}\sqrt{\sum_{i=1}^{n} w_i(y_i - \bar{y}_w)^2}}$$

where $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ and $\bar{y}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$.

This new correlation coefficient r_W can be viewed as the weighted version of Pearson's product moment correlation coefficient which considers only the "good" data points and is expected to have the same maximal breakdown point ($\epsilon* = 50\%$) as the WLS regression.

For our study in this project, while defining w_i as mentioned above, we have used

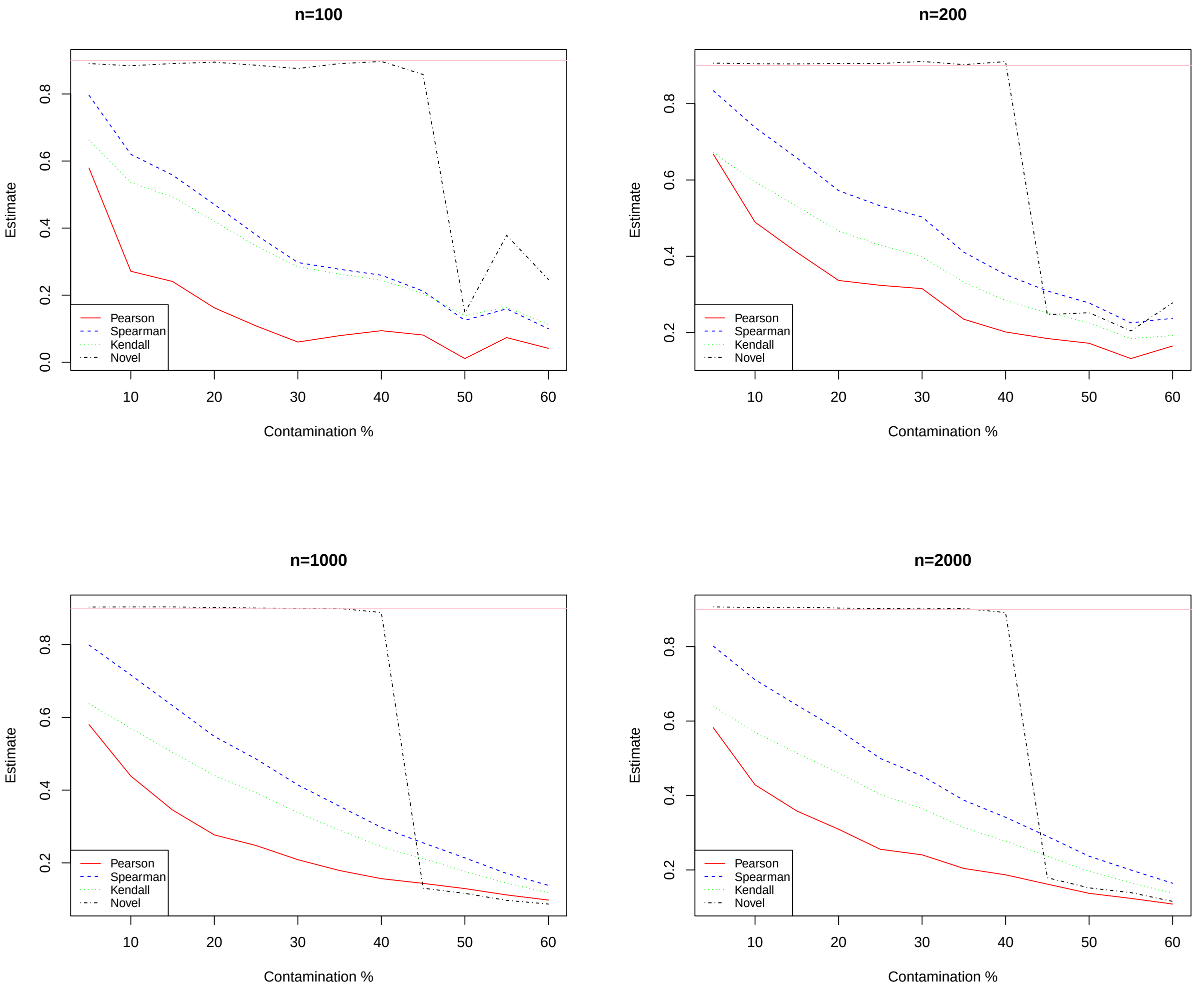
$$\hat{\sigma} = \frac{med_i|e_i - med_j(e_j)|}{0.6745}$$

as an estimate of σ, where e_i denotes the residuals when y is regressed on x.

3 Simulation Results

3.1 Comparison between different measures

Here, we plot the values of the three well known correlation coefficients along with the value of the new correlation coefficient on the same graph and see their behaviour as we increase the amount of contamination in the data. We have taken the true value of the correlation as 0.9 in our study (in the original paper the author studied with correlation coefficient 1 but here we wanted to be more realistic). We do this for different sample sizes.

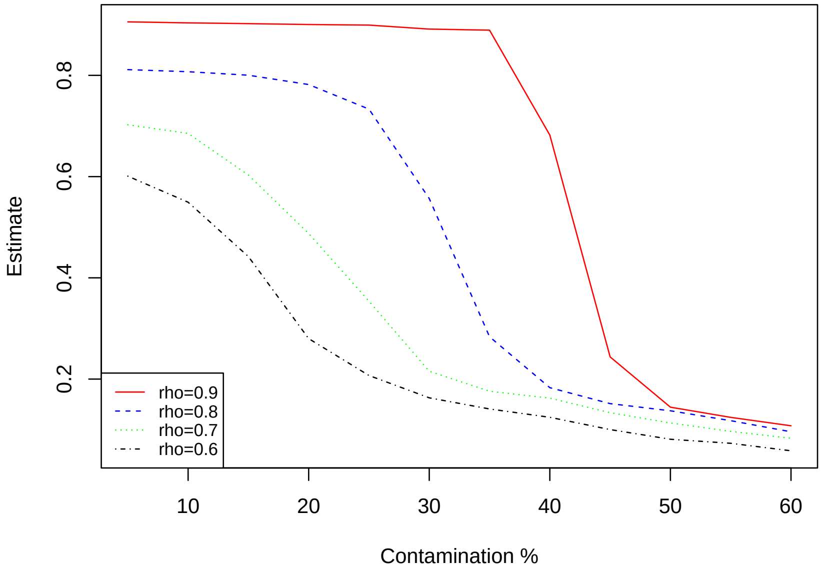


Comparison of breakdown points for different sample sizes

From the graphs, it is easy to see that the newly proposed correlation coefficient has a much higher breakdown point (almost close to 50%) compared to the other measures of correlation.

3.2 Limitations

We have seen that the new proposed correlation coefficient outperforms the conventional ones in terms of robustness. However, in the paper, the simulation study was done for the highest correlation ($\rho = 1$). In my study I worked with a bit different but still a high value ($\rho = 0.95$) of the actual correlation coefficient. A natural question came to my mind whether this new measure would perform well even if the two variables are not highly correlated. Thus, I wanted to check the if the breakdown point changes if we change the value of ρ. I got the following (run over 100 simulations) :



Behaviour of breakdown point for decreasing values of correlation coefficient

It can be observed that as the true value of the correlation between the two variables decrease, the breakdown point of the weighted correlation coefficient also decreases. For negative values of correlation, we observe a similar behaviour i.e. decrease in breakdown point if we increase the true value. In a nutshell, the new correlation coefficient has a very good performance in terms of robustness when the absolute value of the correlation between the two variables is very high. For lower values of correlation, its breakdown point approaches that of the other conventional correlation coefficients.

Also, this correlation coefficient is not symmetric is nature and does not deal with the case of high leverage points (i.e. if there are outliers in the predictor variables).

4 Application on real data

The data we are going to use can be found here in dta format which can be directly imported in R the **read.dta** command.

4.1 Data Description

At first, We use the crime data set that appears in *Statistical Methods and Social Sciences, 3rd Edition* by Agresti and Finlay [AF97] . For a particular state, there are several variables, among which we will use **murders** (per 1,000,000) as our response variable and violent **crimes** (per 1,000,000) and the percentage of the total population that is white (**pctwhite**) as our predictors. We have 51 data points.

Next we use the Star Cluster Data inbuilt in R, which has 47 observations. Here we use log surface temperature of the star as our response variable and log light intensity of the star as our predictor.

4.2 Experiment Results

Crime Dataset : At first we try to visualize how the predictors are related to the response variable.

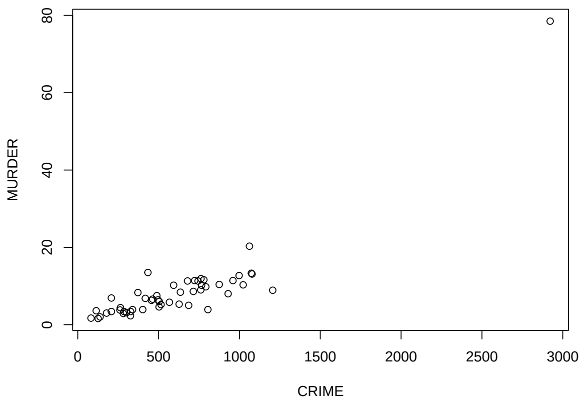


Figure 1: Murder vs Crime

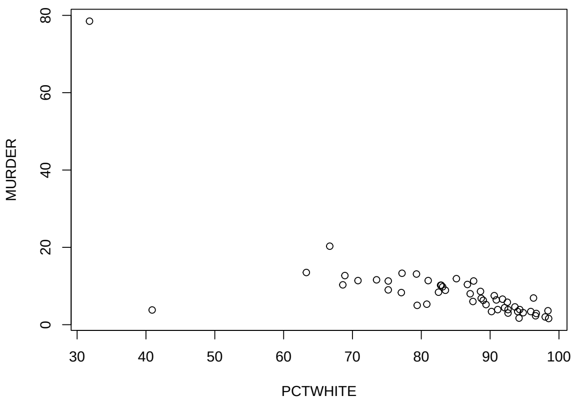


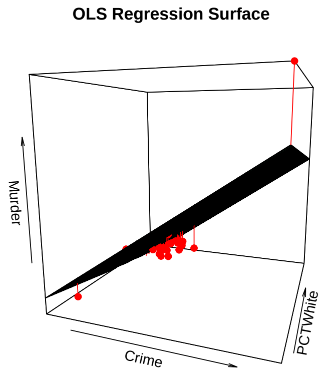
Figure 2: Murder vs % of White Population

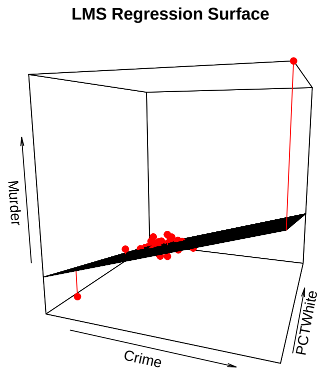
It seems that murder and crime have a very strong positive, almost near perfect linear relationship, which is quite understandable, whereas murder has a strong negative linear relationship with the percentage of white population of the state. So, let us check whether the Pearson's correlation coefficient reflect that. Simultaneously, we also look at the new robust correlation coefficient.

Pair of Variables	r_P	r_W
Murder vs Crime	0.8861963	0.9371983
Murder vs PCTWhite	-0.7061927	-0.9065586

We can observe that, although Pearson's correlation coefficient performs reasonably well in the first case, in the second case it fails to indicate the ”very strong” linear relationship between the variables. The reason is the presence of outliers in the data, which we can see in the above scatterplots. The robust correlation coefficients perform better and reveal the truth better than the Pearson's one.

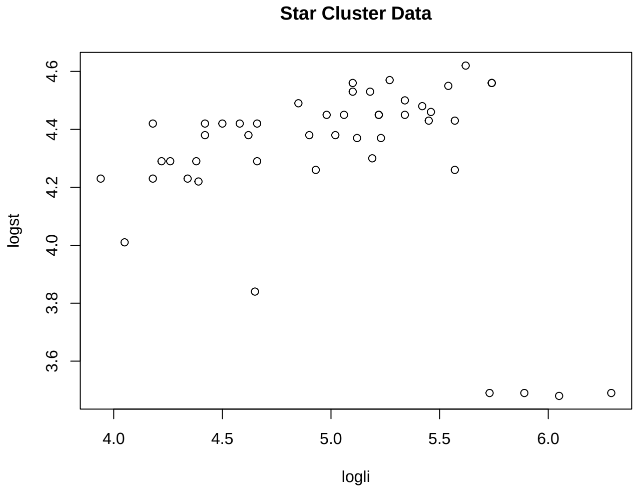
We can also take a look at the 3 dimensional plots of the regression surfaces when ordinary least squares method is applied and when the least median of squares method is applied.


OLS Regression Surface


LMS Regression Surface

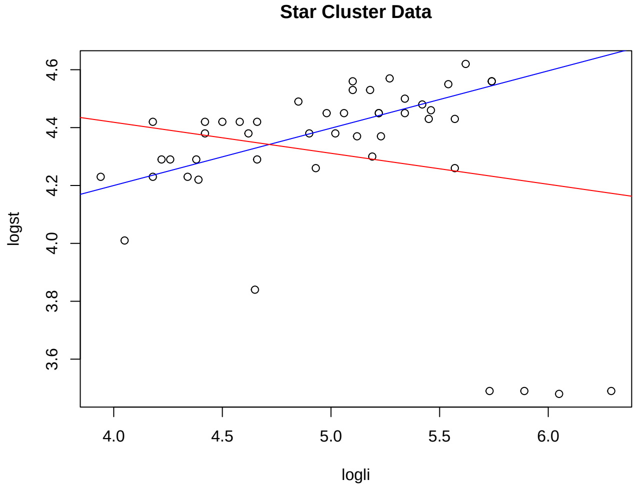
Clearly, it can be seen that the in the case of OLS regression, the outlier pulls the regression surface towards itself, whereas the LMS regression surface fits to the data better, and since the robust correlation coefficient is based on the LMS regression fit, it performs very well. The difference of the performance between these two measures would have been more drastically different if the number of outliers increased.

Star Cluster Dataset : Let us visualize the data at first.


Star Cluster Data

Here we can see there are four outliers, we want to calculate the correlation between the two variables. The

Pearson's product moment correlation coefficient gives value **-0.21** whereas the weighted correlation coefficient based on LMS gives **0.75**. Let us see why does this happen.



In the above picture, the red line is the OLS regression line and the blue line is the LMS regression line. Clearly, the four outliers present in the data pull the regression line so much towards themselves that it reflects negative correlation, which was actually a high positive value in the absence of those outliers. LMS being a much more robust method, the new correlation performs well and is resistant to the effect of those four outliers.

****

References

[Abd90] Mokhtar Bin Abdullah. *On a Robust Correlation Coefficient*. 1990. DOI: 10.2307/2349088. URL: http://dx.doi.org/10.2307/2349088.

[AF97] Alan Agresti and Barbara Finlay. *Statistical methods for the social sciences*. en. 3rd ed. Upper Saddle River, NJ: Pearson, Mar. 1997.

[Rou84] Peter J. Rousseeuw. *Least Median of Squares Regression*. en. Dec. 1984. DOI: 10.1080/01621459.1984.10477105. URL: http://dx.doi.org/10.1080/01621459.1984.10477105.