"Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond": Comment

David E. Allen[a,*], and Michael McAleer[b]

EI2018-33

[a] *School of Mathematics and Statistics, University of Sydney, Department of Finance, Asia University, Taiwan, and School of Business and Law, Edith Cowan University, Australia*
[b] *Department of Finance, College of Management, Asia University, Taiwan, Discipline of Business Analytics, University of Sydney Business School, Australia, Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, The Netherlands, Department of Economic Analysis and ICAE, Complutense University of Madrid, Spain, Department of Mathematics and Statistics, University of Canterbury, New Zealand, and Institute of Advanced Sciences, Yokohama National University, Japan*

Abstract

This note comments on the Generalised Measure of Correlation (GMC) suggested by Zheng et al. (2012). The GMC concept was largely anticipated in a publication 115 years earlier, undertaken by Yule (1897), in the proceedings of the Royal Society. The note is directed at giving Yule (1897) credit for covering the foundations of the topic comprehensively.

Keywords: Skewed correlation, Bravais formula, Generalised Measure of Correlation, Nonlinearity
JEL Codes: C12, C100, C130.

1. Introduction

Zheng et al. (2012) suggest that Pearson's correlation, when used as a measure of explained variance, is well understood, but a major limitation is that it does not account for asymmetry. Zheng et al. (2012) present what they suggest is a broadly applicable correlation measure, and consider a pair of generalized measures of correlation (GMC) that deal with asymmetry in the explained variance, and linear or nonlinear relations between random variables. The authors present examples under which the paired measures are identical, and become a symmetric correlation measure that is the same as the squared Pearson's correlation coefficient, so that Pearson's correlation is a special case of GMC. Zheng et al. (2012) suggest that the theoretical properties of GMC show that GMC can be applicable in numerous applications, and can lead to more meaningful conclusions and improved decision making.

*Corresponding author
Email address: profallen2007@gmail.com (David E. Allen)

Vinod (2015) applied the GMC metric in an economic paper which featured an analysis of development economics markets in a study of 198 countries, and also developed the R library package 'generalCorr' (2017). Allen and Hooper (2018) used the metric to analyse causal relations between the VIX, S&P500, and the realised volatility (RV) of the S&P500 sampled at 5-minute intervals.

Zheng et al. (2012) suggest that they intended to introduce effective and broadly applicable statistical tools for dealing with asymmetry and nonlinear correlations between random variables. For simplicity of illustration, they regard "linear" or "symmetric" as special cases of "nonlinear" or "asymmetric", respectively. In the case of "linear and symmetric," Pearson's correlation coefficient is an extremely important and widely used analytical tool in statistical data analysis. Zheng et al. (2012) claim that 'New dependence measures' that comprise Pearson's correlation coefficient as a special case should be of the greatest interest to practitioners.

The paper is intended to draw attention to the fact that many of the issues addressed by Zheng et al. (2012) were previously anticipated and developed, in the *Proceedings of the Royal Society,* by the British Statistician Udney Yule in (1897), some 115 years earlier! This note sets out Yule's approach and gives Yule (1897) credit for covering the foundations of the topic comprehensively.

2. Yule's (1897) approach to general correlation

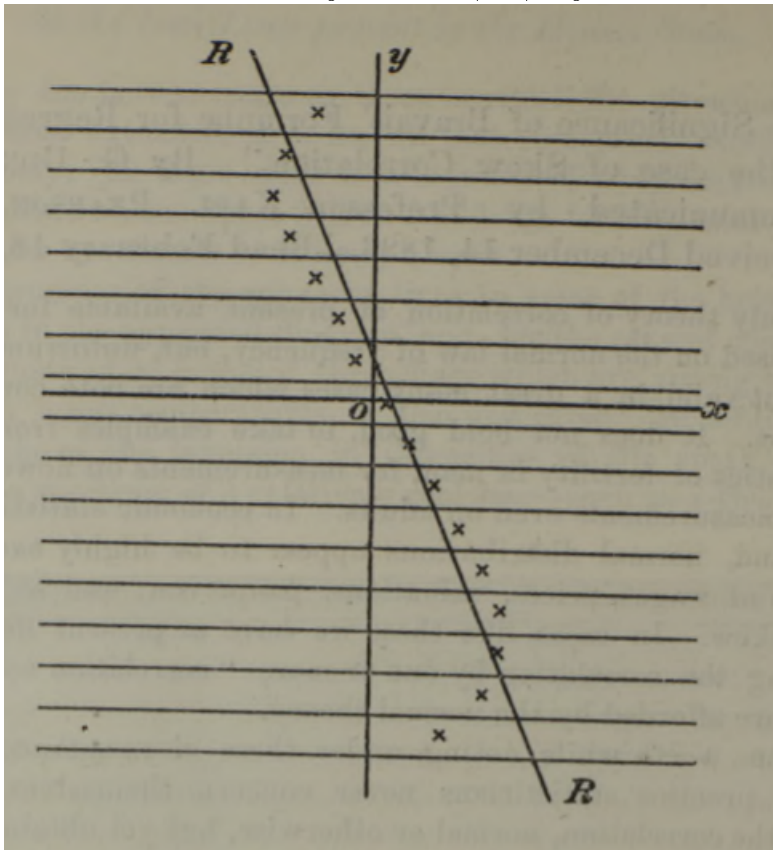
Yule (1897, p.477) observed that: "The only theory of correlation at present available for practical use is based on the normal law of frequency, but, unfortunately, this law is not valid in a great many cases which are both common and important....in economic statistics, on the other hand, normal distributions appear to be highly exceptional: variation of wages, prices, valuations, pauperism, and so forth, are always skew."

He suggests letting Ox and Oy be the axes of a three-dimensional frequency surface drawn through the mean 0 of the surface parallel to the axes of measurement, and the points marked (x) be the means of successive x arrays, lying on some curve that may be called the curve of regression of x on y. Then a line, RR, is fitted to this curve, as shown in Figure 1 (taken from his paper).

In commenting on the diagram, Yule notes that, if the slope of the line RR is positive, large values of x are associated with large values of y, while if negative, large values of x are associated with small values of y.

More importantly, for current purposes, Yule also notes that if the means of the arrays actually lie in a straight line (as in normal correlation), RR must be the equation to the line of the regression. Yule then lets n be the number of observations in any x array, and d be the horizontal distance of the mean of this array from the line RR. He then proposes to subject the line to the condition that the sum of all quantities like nd should be a minimum. In effect, he chooses to use the condition of least squares. He cautions that he does this solely for convenience of the analysis, and that he does not claim any advantages with regard to the probability of the results. He cautions that it would be absurd

Figure 1: Yule's (1897) diagram



to do so, as it is postulated at the outset that the curve of regression is only exceptionally a straight line, so that there can be no meaning in seeking the most probable straight line to represent the regression.

Yule proceeds by letting x and y be a pair of associated deviations, lets σ be the standard deviation of any array about its mean, and writes the equation of a straight line for RR as:

$$X = a + bY.$$

It follows that, for any one array:

$$S\{x - (a + by)\}^2 = S\{x - (a + bY\}^2 = n\sigma^2 + nd^2.$$

Then he extends the meaning of S to sum over the whole surface:

$$S(nd^2) = S\{x - (a + by)\}^2 - Sn\sigma^2,$$

where $Sn\sigma^2$ is independent of a, and is what he terms a characteristic of the surface. It follows that, if $S(nd^2)$ is set to a minimum, this is equivalent to making:

$$S\{x - (a + by)\}^2$$

a minimum. Yule suggests forming a single-valued relation:

$$x = a + by$$

between a pair of associated deviations, such that the sum of squares of errors in estimating any one x from the corresponding y is a minimum. This is simply the line of the regression RR. There will be two such equations to be formed corresponding to the two lines of the regression.

Yule then considers multiple combinations of variables that can be considered as two variables. As x and y represent deviations from their respective means, Yule suggests, using S to denote summation over the whole surface:

$$S(x) = S(y) = 0.$$

The characteristic or regression equations are of the form:

$$\begin{aligned} x &= a_1 + b_1 y \\ y &= a_2 + b_2 x. \end{aligned} \tag{1}$$

Taking the equation for x, the normal equations for a_1 and b_1 are:

$$\begin{aligned} S(x) &= N(a_1) + b_1 S(y) \\ S(xy) &= a_1 S(y) + b_1 (Sy^2) \end{aligned} \tag{2}$$

with N being the number of correlated pairs. The first equation gives:

$$a_1 = 0$$

while the second gives:

$$b_1 = \frac{S(xy)}{S(y^2)}.$$

Yule then lets $S(x^2) = N(\sigma_1^2)$, $S(y^2) = N(\sigma_2^2)$, and $S(x, y) = Nr\sigma_1\sigma_2$, where $\sigma_1$and σ_2 are the two standard deviations or mean square errors, and r is Bravais' (1846) value of the coefficient of correlation. Yule rewrites b_1 as:

$$b_1 = r\frac{\sigma_1}{\sigma_2}. \tag{3}$$

Similarly, when $a_2 = 0$:

$$b_2 = r\frac{\sigma_2}{\sigma_1}. \tag{4}$$

The expressions on the right of (3) and (4) are the values obtained by Bravais on the assumption of normal correlation for the regressions of x on y, and of y on x. Therefore, the Bravais values for the regressions are simply the values of b_1 and b_2 that make:

$$S(x - b_1 y)^2 \quad and \quad S(x - b_2 y)^2$$

their respective minima.

Denis (2000) observes that Bravais (1846) mathematically found the equation of the normal surface for the frequency of error. Using both analytic and geometric methods, Bravais also essentially found what would eventually be coined the "regression line", by investigating how the various elliptical areas of the frequency surface vary according to observed quantities. However, astronomers of the time were far more interested in "disposing" of this common error variance, largely due to the concern that errors would multiply, not compensate, when combining celestial observations.

Yule (1897) suggests proceeding by letting n be the number of correlated pairs in any one array taken parallel to the axis of x, and θ be the angle that the line of regression makes with the axis of y. For a single array:

$$S(xy) = yS(x) = ny^2 tan\theta,$$

or extending S to summation over the whole surface:

$$S(xy) = Ntan\theta\sigma_2^2,$$

or:

$$tan\theta = r\frac{\sigma_1}{\sigma_2}.$$

If the regression is linear, Bravais's formula may be used without investigating the normality of the distribution.

In the general case, both coefficients of regression must have the same sign, namely, the sign of r. Hence, either regression will serve to indicate whether there is correlation or not. Yule suggests that the regressions are not convenient measures of correlation, as it may be found that:

$$b_1 > b_1^{'}, \quad b_2 < b_2^{'},$$

where $b_1, b_2$and $b_1^{'} b_2^{'}$ are the regressions in the two cases. Yule queries to which distribution should we attribute the greater correlation? He observes that Bravais' coefficient solves the difficulty by taking the geometrical mean of the two regressions as the measure of correlation. It will still remain valid for non-normal correlations.

Yule generalises the argument by suggesting that, instead of measuring x and y in arbitrary units, each is measured in terms of its own standard deviation:

$$\frac{x}{\sigma_1} = \rho \frac{y}{\sigma_2} \tag{5}$$

and solve for ρ by the method of least squares. A constant on the right-hand side can be ignored, as it would vanish, yielding:

$$\rho = \frac{S(xy)}{S(y^2)} \frac{\sigma_2}{\sigma_1} = r. \tag{6}$$

If measured x and y are each in terms of their respective standard deviations, r becomes the regression of x on y, and the regression of y on x.

Forming the sums of squares of the residuals in equations (1) and (6), and inserting the values of b_1, b_2, and ρ, gives:

$$\begin{gathered} S(x - b_1)^2 = N\sigma_1^2(1 - r^2) \\ S(x - b_2)^2 = N\sigma_2^2(1 - r^2) \\ S(\tfrac{x}{\sigma_1} - \rho\tfrac{y}{\sigma_2})^2 = S(\tfrac{y}{\sigma_2} - \rho\tfrac{x}{\sigma_1})^2 = N(1 - r^2), \end{gathered} \tag{7}$$

each of which is positive. Hence, r cannot be greater than unity. If r is equal to unity, each of the above becomes zero.

However,

$$S(\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2})^2$$

can only vanish if:

$$\frac{x}{\sigma_1} \pm \frac{y}{\sigma_2} = 0$$

in every case, or if the following relation holds:

$$\frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3} = ... = \pm\frac{\sigma_1}{\sigma_2}, \tag{8}$$

with the sign of the last term in (7) dependent on the sign of r. Hence, the statement that two variables are perfectly correlated implies that relation (7) holds, or that all pairs of deviations bear the same ratio to one another. It follows that, where the means of the arrays are not collinear, or the deviation of the means of the arrays is not a linear function of the deviation, r cannot be unity. If the regression model is far from linear, caution must be used in using r to compare two different distributions.

3. Limitations inYule (1897)

Zeng et al. (2012) point out that, despite its ubiquity, there are inherent limitations in the Pearson correlation coefficient when it is used as a measure of dependency. One limitation is that it does not account for asymmetry in the explained variances, which are often innate among nonlinearly dependent random variables. As a result, measures dealing with asymmetries are needed.

In order to meet this requirement, they developed Generalized Measures of Correlation (GMC). They commence with the familiar linear regression model, and the partitioning of the variance into explained and unexplained components.

$$Var\,(X) = Var(E(X \mid Y) + E(Var(X \mid Y)), \tag{9}$$

whenever $E(Y^2) < \infty$ and $E(X^2) < \infty$. Note that $E(Var(X \mid Y))$ is the expected conditional variance of X given Y, and therefore $E(Var(X \mid Y)/Var(X)$ can be interpreted as the explained variance of X by Y. Thus, we can write:

$$\frac{E(Var(X \mid Y))}{Var(X)} = 1 - \frac{E(Var(X \mid Y))}{Var(X)} = 1 - \frac{E(\{X - E(X \mid Y)\}^2}{Var(X)}.$$

The explained variance of Y given X can be defined similarly. This leads Zheng et al. (2012) to define a pair of generalised measures of correlation (GMC) as:

$$\{GMC(Y \mid X),\, GMC(X \mid Y)\} = \left\{1 - \frac{E(\{Y - E(Y \mid X)\}^2}{Var(Y)},\; 1 - \frac{E(\{X - E(X \mid Y)\}^2}{Var(X)}\right\}. \tag{10}$$

This pair of GMC measures has some attractive properties. It should be noted that the two measures are identical when $(X,\, Y)$ is a bivariate normal random vector.

4. Conclusion

Zheng et al. (2012) provide a convincing explanation of the properties of the measure they refer to as a generalised meaure of correlation (GMC). This note draws attention to the fact that some of the properties of their suggested metric were previously explored by Yule (1897) in analysing skew correlation 115 years earlier.

References

[1] Allen, D.E and V. Hooper (2018) Generalized Correlation Measures of Causality and Forecasts of the VIX using Non-linear Models, *Sustainability*, 10(8: 2695), 1-15.

[2] Bravais, A. (1846), Analyse mathématique sur les probabilités des erreurs de situation d'un point, Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France, 9, 255-332.

[3] Denis, D.J. (2000), The Origins of Correlation and Regression: Francis Galton or Auguste Bravais and the Error Theorists?, Paper presented at the 61st Annual Convention of the Canadian Psychological Association, Ottawa, Canada, 29 June, 2000.

[4] Vinod, H.D. (2015), Generalized Correlation and Kernel Causality with Applications in Development Economics, *Communications in Statistics - Simulation and Computation*, accepted Nov. 10, 2015, URL http://dx.doi.org/10.1080/

[5] Yule, G.U. (1897), On the Significance of Bravais Formulæ for Regression, in the case of skew correlation, *Proceedings of The Royal Society London*, 477-489.

[6] Zheng, S., N-S, Shi and Z. Zhang (2012), Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond, *Journal of the American Statistical Association*, 107, 1239-1252.