



Measuring Triangular Macro Correlation¹

Othmar W. Winker

Professor Emeritus

ABSTRACT

This study explores a new and simpler way of assessing the co-relation of the interaction between variables having opposite effects like “If one grows the other declines”. The proposed Triangular correlation treats these data as an aggregate or single bloc of data, situated within the three sides of a triangle formed by the hypotenuse, the adjacent side of the triangle, on the horizontal axis, and the opposing side of that triangle, on the vertical axis. The hypotenuse is the longest side of the triangle, a line fitted by least square to the highest data. The intersection of the hypotenuse with the axes determines the length of both segments and their ratio, the mega-correlation coefficient M , reflects the data accumulation inside the triangle. M is interpreted like r^2 , even though these coefficients measure different, complementary aspects of the co-relation.

Key words: regression analysis, correlation, assessing the interaction between two variables

BACKGROUND

The English biometricians Galton and Pearson’s experience with measures of the height of fathers matched with the height of their sons led to their discovery that the sons’ height “regressed” to that of their fathers. They formulated mathematically that **co-relation as correlation**, the trend of their data as the **regression line**, complementing it with the ‘**coefficient of correlation**’. Since then it has been tacitly assumed that two variables generally develop in the same direction: when one is higher, the other usually is also higher.

As a young instructor² I tried correlation analysis with my first 115 students’ semester grades³ as their Y-values and the percentage of classes missed, as their X-values. Their data, Figure 3, showed a general drift from high grades in the upper left, of students earning higher grades who missed few classes, to students on the lower right earning lower grades who missed many classes. This reverse relation of the two variables confirmed the intuition that class attendance and grades were connected. That was obvious because textbooks in my students’ language, Spanish, were not available, attending lectures and taking notes were the conditions of passing that course, but did not guarantee good grades.

¹ This presentation and the name “Triangular Macro Correlation” is based on and treated following essentially the article “On Triangular Correlation,” Othmar W. Winkler in the “Proceedings of the Business and Economic Statistics-Section of the ASA”, Alexandria, VA 1980. This study is further based on “chapter 9.4 “Different Forms of Data Association in the Social Sciences”, in “Interpreting Economic and Social Data – A Foundation of Descriptive Statistics,” Springer, 2009, Othmar W. Winkler

² In 1949/50 the author was invited to teach economic statistics in the newly expanded economics department of the Universidad Central de Venezuela, in Caracas,

³ The established grading system then was from “20,” excellent, comparable to A+ or A, to “10” just barely passing, a “D”, below that, a 9 or less, failure, F“

The data appeared to form a triangle, the adjacent side of that triangle on X, the opposite side of the triangle on Y and a line fitted to the highest points, as the hypotenuse. A search for a sample of Triangular Correlation for a similar constellation of data in other disciplines discovered the following graphs⁴.

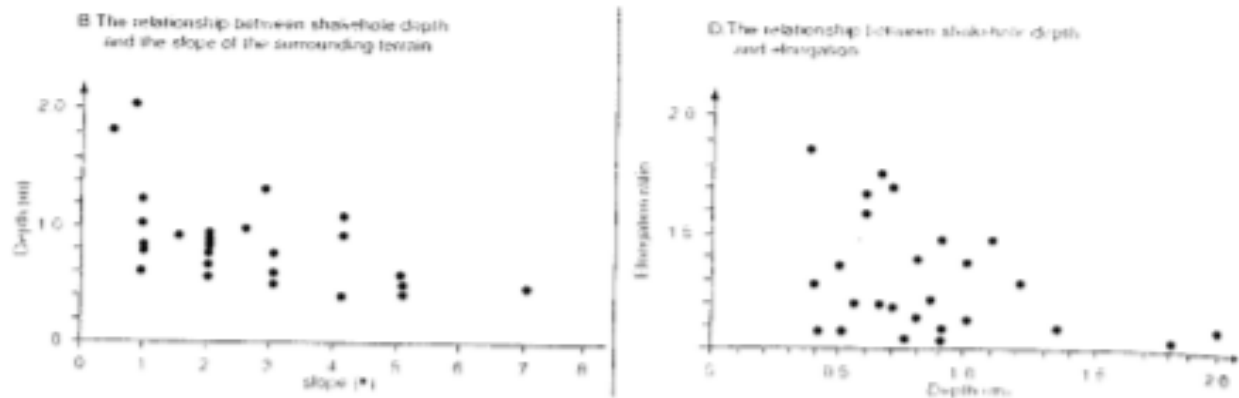


Figure 1

Size, Depth and slope of shake-holes

That triangular data plot suggested another, different approach: instead of using the 115 individual student records, treating that triangular group of records as an aggregate, a unit. The behavior of a large flock of birds come to mind, e.g. starlings, pigeons or crows, that fly and alight together with a few stray-birds flying outside, along the flock. That view of socio-economic statistics, mostly dealing with aggregate data, differs from the demand of exact measurements in biostatistics and the sciences. Relying on groups instead of individual data is like viewing the economy through a macroeconomic lens, studying the big overarching features of an economy.

The regression line of grades and class attendance of these 115 students, according to the accepted formula since Galton-Pearson, was $Y_{115} = 15.755247 - .0634721X$, the coefficient of correlation $r = -0.4285019$ and $r^2 = .1836138$. This regression line had a negative slope, as expected, but was nearly horizontal, intersecting with the X-axis at $X = 248.22\%$ - notice, the presentation of the horizontal axis in Figure 3 goes only to 53%! The intuitively more informative r^2 recognizes an existing but disappointingly weak relationship between the variables student grades and class attendance.

A STATISTICAL MACRO CORRELATION APPROACH

The graphs of that situation, **Figures 1** and **3**, resemble rectangular triangles, formed by its adjacent and opposite sides on X and Y, and the hypotenuse, formed by a least-square line fitted through the 19 highest data points as $Y_{19} = 20.161351 - 0.1626217X$. In an earlier publication I called this “Triangular correlation”. These 19 student records were neither a random nor a representative sample of the 115 student records. They were, like the outliers of a large flying

flock of starlings or pigeons, forming the upper limit of the data in triangular shape. The intersections of that fitted equation with the Y axis was 20.16 and that with the X- axis

⁴ Taken from “Shake holes” a morphometric field project for sixth-form geographers. In *Geography*, vol. 65, pt. 3, July 1980, No. 288. A similar search on Google brought up ten publications in medical/cancer research who responded had the words **Triangular Correlation** headlined. All were published at or after 2000. The author will honor your request for sending that collection by email.

was 124.5. Most of the data were located within the triangle formed by the hypotenuse, the adjacent and the opposite sides.

The scales on both axes had to be equal before computing the M ratio. No change was needed for the percentage-scale on X, ‘absence from class’. But the scale on Y, of the grades 9 to 20, had to be converted to a percentage from 0% to 50%. That was accomplished by subtracting 9 from each vertical scale value, then dividing this difference by 11, the width of this grading schedule, and that ratio multiplied by 50, to match the high end of the scale on X. $(x-9)/11 \cdot 50$. This conversion changes all existing Y values to numbers between 0 and 50. It also changes the equation of the hypotenuse to the converted values $Y_{19} = 48.849312 - 0.6913888 X$. See Figure 3 (the blue line).

As the adjacent side of the triangle, “B”, see Figure 2, usually is longer than the opposite side “A”, it is advantageous to use this larger side as the denominator for $M = 48.855/69.84609 = .699468$ facilitating comparison and interpretation of the results of both regressions $r^2 \sim 18\%$ but $M \sim 70\%$. For these 115 observations - they are not measurements but administrative records - these two correlations indicate the influence of class attendance on grade, but not grade on class attendance as the two are successive, not simultaneous event variables. In this situation M is a stronger, better suited measure for these contrary related variables than the usual coefficient of correlation r^2 . Compare $M = 70\%$ to $r^2 = 18\%$, how M also responds and aligns much better with the actual experience of teachers.

What do these two quite different assessments of these data in a triangular setting mean apart from M’s relative simplicity, ease and speed? Besides these advantages of the macro-statistical approach, the classical correlation, very importantly, does not appear appropriate for this kind of contra-related variable.

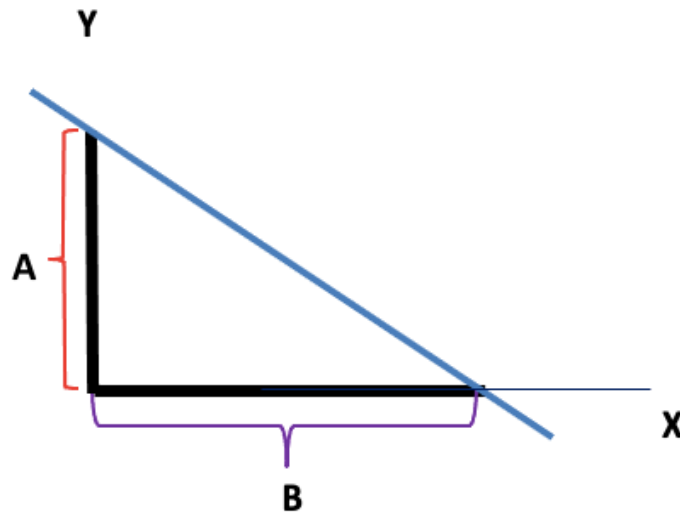


Figure 2

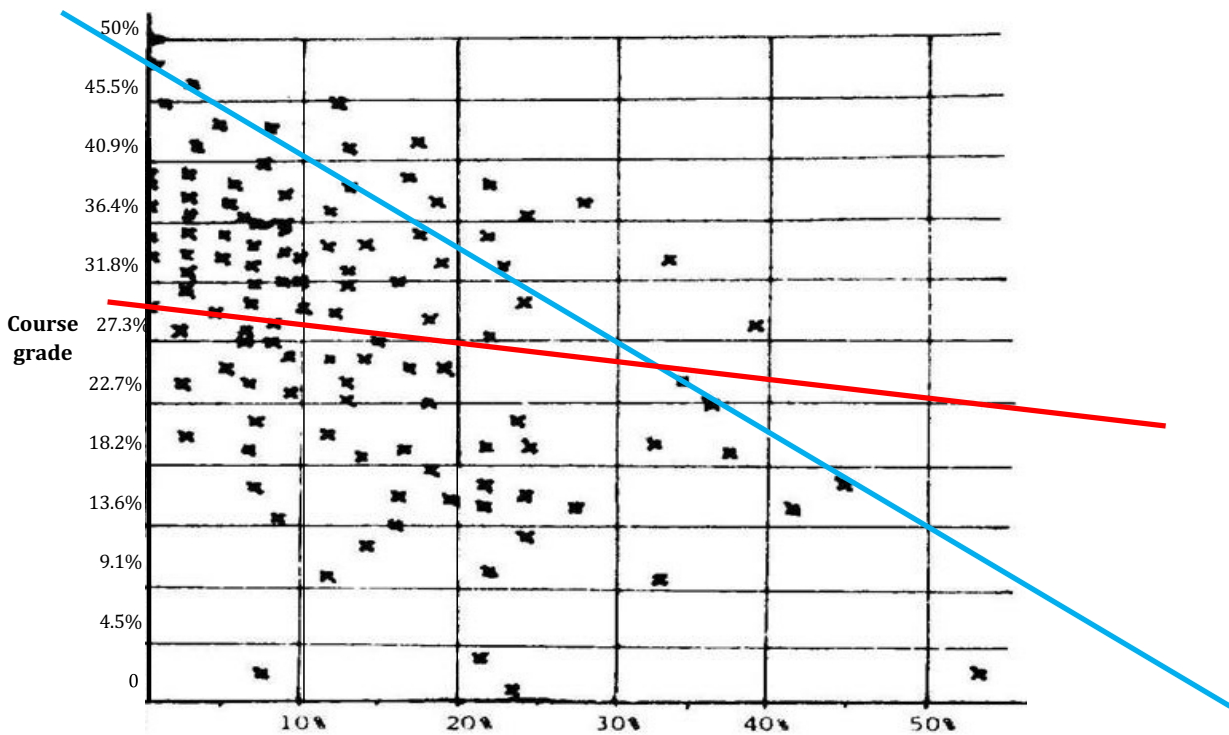


Figure 3

Percent of Classes missed

Red line: Least square fit to 115 data points (the different Y-scale not shown) Blue Line: $Y_{19} = 48.849312 - 0.6913888 X$ least square fit to 19 highest points

SOME CONCLUSIONS

The standard regression analysis first determines the regression line of all the data to reveal the general direction of the association and effect of the two variables. Then the squared distance of each record from that regression line is determined. The actual data then are

considered to be only “deviations”. The coefficient ‘r’ makes its assessment of the strength of the influence of the variables through the size of the sum of squares of the deviations from the regression line.

The macro approach M, aims at the big picture. It does that without the need for a regression line – the line fitted as the hypotenuse is something quite different - as a center to judge the distribution of the individual data. The ratio M is not based on how the data are located inside the triangle but is taken as an indirect indication of the data distribution within the triangle. It treats each datum as an independent piece of the reality of the interaction of the variables, not as an error or a ‘deviation’ from some imaginary center of the interaction between the variables. M maxes out at 1.0 if both sides of the triangle are equally long, a state of the perfect, balanced interaction between the investigated variables. It does not show how that happened as there are various arrangements of the data within the triangle to yield $M = 1$ to occur. That happens when the data within are equally and symmetrically aligned along either of the sides of that triangle without paying attention to the detailed distribution of the record (the data) and ignores the standard regression line.

That is quite different with r and r^2 : The farther away the data are from that regression equation, as their idealized center of gravity, the lower is the value of r^2 . M, to the contrary, works without use of the individual records – except for the few highest data on the upper border, to place the hypotenuse. By concentration on the big picture, M captures the overall relation between the variables at the cost of paying no attention to the detailed distribution of the data within that triangle. The ratio M approaches 0 or infinity the closer the data are aligned with the sides of the triangle along Y or X. In that case M indicates ‘barely or no interaction between the two variables and the data tend to be aligned as a one-dimensional frequency distribution of one or the other of the two variables.

To get a fuller understanding of a triangular distribution, both the standard r^2 and the faster M correlation coefficient, based on different assumptions and procedures, would reveal the effects of the variables on each other. If time is not an issue to perform this statistical analysis, both procedures would reveal the common as well as different aspects of your data.