

The Correlation Coefficient: Its Values Range Between Plus/Minus 1, or Do They? Bruce Ratner, Ph.D.

The “correlation coefficient” was coined by Karl Pearson in 1896. Accordingly, this statistic is over a century old, and is still going strong. It is one of the most used statistics today, second to the mean. The correlation coefficient’s weaknesses and warnings of misuse are well documented. As a fifteen-year practiced consulting statistician, who also teaches statisticians continuing and professional studies for the Database Marketing/Data Mining Industry, I see too often the weaknesses and warnings are not heeded. Among the weaknesses/uses, there is one that is rarely mentioned: the correlation coefficient interval $[-1, +1]$ is restricted by the individual distributions of the two variables being correlated. The purpose of this article is: 1) to introduce the affects the distributions of the two individual variables have on the correlation coefficient interval; and 2) thusly, to provide a procedure for calculating an *adjusted correlation coefficient*, whose realized correlation coefficient interval is often shorter than the original one.

Basics of the Correlation Coefficient

The correlation coefficient, denoted by r , is a measure of the strength of the straight-line or linear relationship between two variables. The well-known correlation coefficient is often misused because its linearity assumption is not tested. The correlation coefficient can – by definition, i.e., theoretically – assume any value in the interval between $+1$ and -1 , including the end values plus/minus 1.

The following points are the accepted guidelines for interpreting the correlation coefficient:

1. 0 indicates no linear relationship.
2. $+1$ indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
3. -1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
4. Values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) linear relationship via a shaky linear rule.
5. Values between 0.3 and 0.7 (0.3 and -0.7) indicate a moderate positive (negative) linear relationship via a fuzzy-firm linear rule.
6. Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong positive (negative) linear relationship via a firm linear rule.
7. The value of r squared, called the coefficient of determination, and denoted R-squared, is typically interpreted as “the percent of variation in one variable explained by the other variable,” or “the percent of variation shared between the two variables.”
Good things to know about R-squared:
 - a) R-squared is the correlation coefficient is the between the observed and modeled (predicted) data values.
 - b) R-squared can increase as the number of predictor variables in the model increases; R-squared does not decrease. Modelers unwittingly may think a “better” model is being built, as s/he has a tendency to include more (unnecessary) predictor variables in the model. Accordingly, an adjustment of R-squared was developed, appropriately called adjusted R-squared. The explanation of this statistic is the same as R-squared, but it penalizes the statistic when unnecessary variables are included in the model.
 - c) Specifically, the adjusted R-squared adjusts the R-squared for the sample size and the number of variables in the regression model. Therefore, the adjusted R-squared allows for an “apples-to-apples” comparison between models with different numbers of variables and different sample sizes. Unlike R-squared, adjusted R-squared does not necessarily increase if a predictor variable is added to a model.
 - d) R-squared is a first-blush indicator of a good model. R-squared is often misused as the measure to assess which model produces better predictions. The RMSE (root mean squared error) is the measure for determining the better model. The smaller the RMSE value, the better the model, viz., the more precise the predictions. It is usually best to report the RMSE rather than mean squared error (MSE), because the RMSE is measured in the same units as the data, rather than in squared units, and is representative of the size of a “typical” error. The root mean squared error is a valid indicator of relative model quality only if it is well-fitted, e.g., if the model is neither overfitted nor underfitted.
8. Linearity Assumption: The correlation coefficient requires that the underlying relationship between the two variables under consideration is linear. If the relationship is known to be linear, or the observed pattern between the two variables appears to be linear, then the correlation coefficient provides a reliable measure of the strength of the linear relationship. If the relationship is known to be nonlinear, or the observed pattern appears to be nonlinear, then the correlation coefficient is not useful, or at least questionable.

Calculation of the Correlation Coefficient

The calculation of the correlation coefficient for two variables, say X and Y, is simple to understand. Let zX and zY be the standardized versions of X and Y, respectively. That is, zX and zY are both re-expressed to have means equal to zero, and standard deviations (std) equal to one. The re-expressions used to obtain the standardized scores are in equations (1) and (2):

$$zX_i = [X_i - \text{mean}(X)] / \text{std}(X) \tag{1}$$

$$zY_i = [Y_i - \text{mean}(Y)] / \text{std}(Y) \tag{2}$$

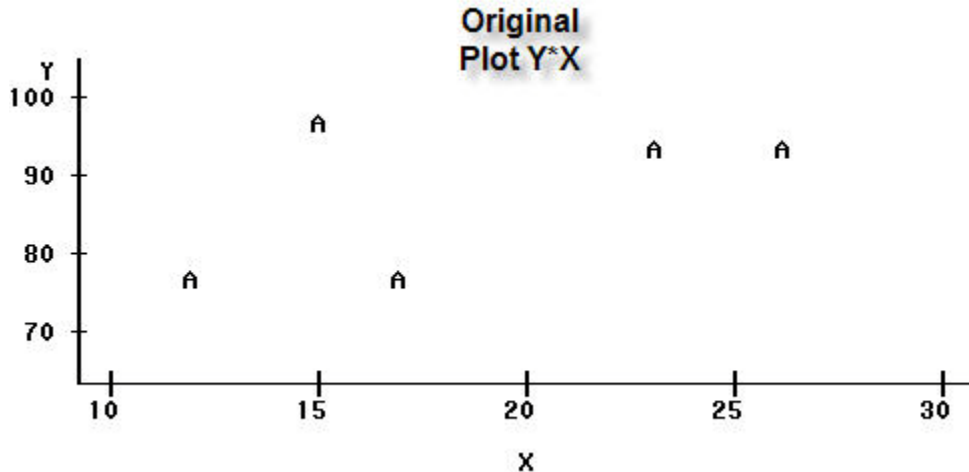
The correlation coefficient is defined as the mean product of the paired standardized scores (zX_i, zY_i) as expressed in equation (3).

$$r_{X,Y} = \text{sum of } [zX_i * zY_i] / (n-1), \text{ where } n \text{ is the sample size} \tag{3}$$

For a simple illustration of the calculation, consider the sample of five observations in Table 1. Columns zX and zY contain the standardized scores of X and Y, respectively. The last column is the product of the paired standardized scores. The sum of these scores is 1.83. The mean of these scores (using the adjusted divisor n-1, not n) is 0.46. Thus, **r_{X,Y} = 0.46**.

Table 1.					
Calculation of Correlation Coefficient					
obs	X	Y	zX	zY	zX*zY
1	12	77	-1.14	-0.96	1.11
2	15	98	-0.62	1.07	-0.66
3	17	75	-0.27	-1.16	0.32
4	23	93	0.76	0.58	0.44
5	26	92	1.28	0.48	0.62
mean	18.6	87.0	sum = 1.83		
std	5.77	10.32			
n = 5			r = 0.46		

For sake of completeness, I provide the plot of the original data, Plot Y and X, below. Unfortunately, the small sample size renders the plot visually unhelpful.



Rematching

As mentioned above, the correlation coefficient theoretically assumes values in the interval between +1 and -1, including the end values plus/minus 1. (An interval that includes the end values is called a closed interval, and is denoted with left and right square brackets: [, and], respectively. Accordingly, the correlation coefficient assumes values in the closed interval [-1, +1].) However, it is not well known that the correlation coefficient closed interval is restricted by the shapes (distributions) of the individual X data, and the individual Y data. The extent to which the shapes of the individual X and individual Y data differ affects the length of the realized correlation coefficient closed interval, which is often shorter than the theoretical interval. Clearly, a shorter realized correlation coefficient closed interval necessitates the calculation of the **adjusted correlation coefficient** (to be discussed below).

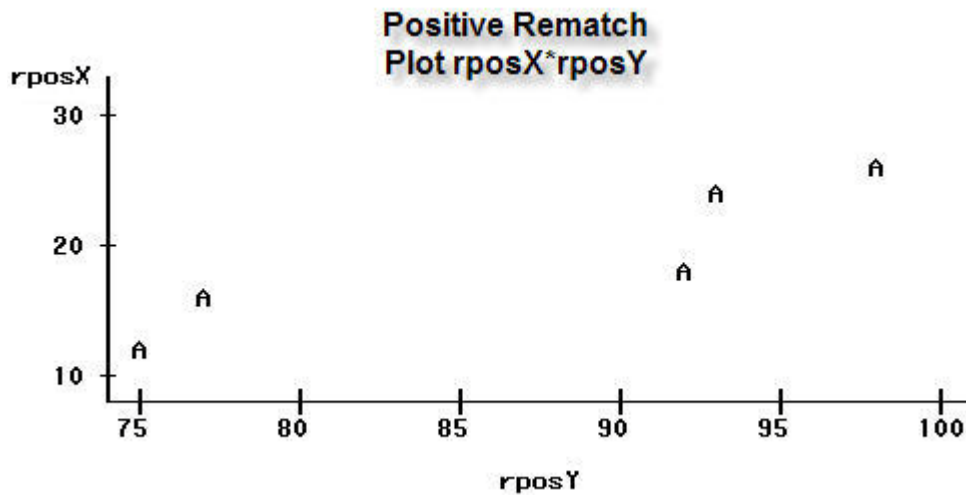
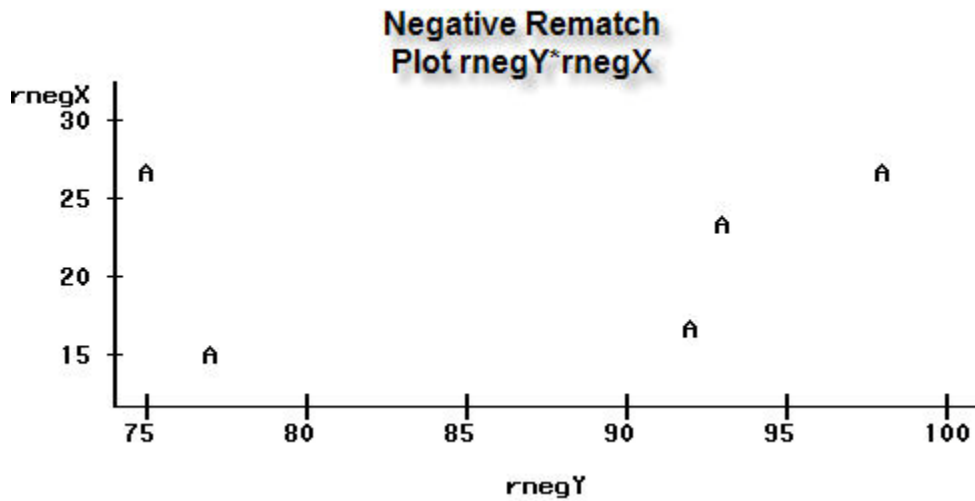
The length of the realized correlation coefficient closed interval is determined by the process of **rematching**. Rematching takes the original (X, Y) paired data to create new (X, Y) "rematched-paired" data **such that** all the rematched-paired data produce the strongest positive and strongest negative relationships. The correlation coefficients of the strongest positive and strongest negative relationships yield the length of the realized correlation coefficient closed interval. The rematching process is as follows:

1. The strongest positive relationship comes about when the highest X value is paired with the highest Y value; the 2nd highest X is paired with the 2nd highest Y value, and so on until the lowest X value is paired with the lowest Y value.
2. The strongest negative relationship comes about when the highest, say, X value is paired with the lowest Y values, the 2nd highest X was paired with the 2nd lowest Y value, and so on until the highest X value is paired with the lowest Y value.

Continuing with the data in Table 1, I rematch the X-Y data in Table 2, below. The rematching produces:

$r_{x,y}$ (negative rematch) = - 0.99 and **$r_{x,y}$ (positive rematch) = + 0.90.**

For sake of completeness, I provide the "rematched" plots. Unfortunately, the small sample size renders the plots visually unhelpful. The plots of the negative and positive rematched data, Plot rnegY and rnegX, and Plot rposY and rposY, respectively, are below.



So, just as there is an adjustment for R-squared, there is an adjustment for the correlation coefficient due to the individual shapes of the X and Y data. Thus, the restricted, realized correlation coefficient closed interval is $[-0.99, +0.90]$, and the adjusted correlation coefficient can now be calculated.

Calculation of the Adjusted Correlation Coefficient

The **adjusted correlation coefficient** is obtained by dividing the original correlation coefficient by the rematched correlation coefficient whose sign is that of the sign of original correlation coefficient. The sign of adjusted correlation coefficient is the sign of original correlation coefficient. If the sign of the original r is negative, then the sign of the adjusted r is negative, even though the arithmetic of dividing two negative numbers yields a positive number. The expression in (4) provides only the numerical value of the adjusted correlation coefficient. In this example, the adjusted correlation coefficient between X and Y is defined in expression (4): the original correlation coefficient with a positive sign is divided by the positive-rematched original correlation.

$$\Gamma_{X,Y}(\text{adjusted}) = \Gamma_{X,Y}(\text{original}) / \Gamma_{X,Y}(\text{positive rematch}) \quad (4)$$

Thus, $\Gamma_{X,Y}(\text{adjusted}) = 0.51 (= 0.46/0.90)$, a 10.9% increase over the original correlation coefficient.

obs	Original (X,Y)		Positive Rematch		Negative Rematch	
	X	Y	X	Y	X	Y
1	12	77	26	98	26	75
2	15	98	23	93	23	77
3	17	75	17	92	17	92
4	23	93	15	77	15	93
5	26	92	12	75	12	98
r	0.46		+0.90		-0.99	

Implication of Rematching

The correlation coefficient is restricted by the observed shapes of the individual X and individual Y values. The shape of the data has the following effects:

1. Regardless of the shape of either variable, symmetric or otherwise, if one variable's shape is different from the other variable's shape, the correlation coefficient is restricted.
2. The restriction is indicated by the rematch.
3. It is not possible to obtain perfect correlation unless the variables have the same shape, symmetric or otherwise.
4. A condition that is necessary for a perfect correlation is the shapes must be the same; but, it does not guarantee a perfect correlation.

Conclusion

The everyday correlation coefficient is still going strong after its introduction over one hundred years ago. The statistic is well studied and its weakness and warnings of misuse unfortunately, at least for this author, have not been heeded. Among the weakness/uses, there is one that is rarely mentioned: The restriction on the values the correlation coefficient assumes; namely, the observed values fall within a shorter than the always taught [-1, 1] interval. I introduce the affects the individual distributions of the two variables on the correlation coefficient closed interval; and, thusly, provide a procedure for calculating an *adjusted correlation coefficient*, whose realized correlation coefficient closed interval is often shorter than the original one, which in turn reflects a more precise measure of a linear relationship between the two variables under study.

(Article: Ratner, B., *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, CRC, 2003)

Related Articles

1. [Genetic Data Mining Method for the Proper Use of the Correlation Coefficient](#)
2. [Calculating the Average Correlation Coefficient](#)
3. [Different Data, Identical Regression Models: Which Model is Better?](#)
4. [Variable Selection Methods in Regression: Many Statisticians Know Them, But Few Know They Produce Poorly Performing Models](#)
5. [Genetic vs. Statistic Regression - A Comparison](#)