

How does one think about a reported correlation coefficient? Some authors report the value of a computed coefficient while others routinely report the squared coefficient, reasoning that this, the “fraction of variance explained” is a better number to plug into thinking about what a coefficient is saying.

Preparing to teach about correlation, I thought about this issue and decided to look into the simplest possible situation for guidance. There are two $N(0,1)$ random variables, e and x , independent of each other. Form a new variable y that is part x and part e and compute the correlation between x and y .

$$y = ax + (1 - a)e \quad 0 \leq a \leq 1$$

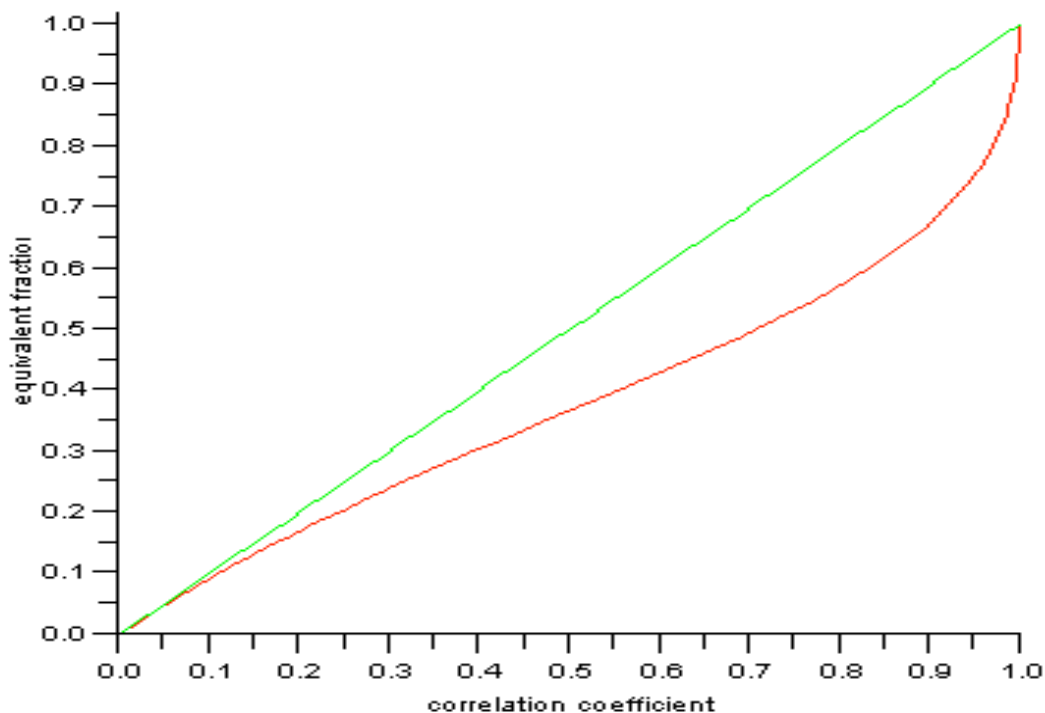
The variable y is now composed of fraction a that is x and a fraction $(1-a)$ that is noise of the same magnitude as x . The correlation between x and y is not a as one might wish but is instead

$$r_{xy} = a / \sqrt{a^2 + (1 - a)^2}$$

The correlation is always larger than a , so it overstates the strength of the relationship using “strength” to mean equivalent fraction. On the other hand, especially for small correlations, a is close. It seems that the familiar ploy of discussing a squared correlation as a proper interpretation is not appropriate.

If for example a correlation of 0.25 is found between two variables, “correct thought” is to think of the relationship as 0.20, the value of a corresponding to a correlation of 0.25, and not 0.06, the squared correlation coefficient.

The figure shows the relationship between correlation, on the horizontal axis, and the equivalent a or fractional composition on the vertical axis.



Imagine that some trait were literally the sum of 30% additive gene effects and 70% random environmental effects. The correlation of genotype and trait would be 0.39 while the squared correlation would be 0.16, seriously understating the “true” value of 0.30. The correlation between the environment e and the trait would be 0.92, seeming high, but this seriously overstates the “true” value of 0.70.

