Correlation and Linearity

The idea of correlation arises naturally for two random variables that have a joint distribution that is bivariate normal. For each individual variable, two parameters a mean and standard deviation are sufficient to fully describe its probability distribution. For the joint distribution, a single additional parameter is required the correlation.

If X and Y have a bivariate normal distribution, the relationship between them is linear: the mean of Y, given X, is a linear function of X ie:

$$E(Y|X) = \alpha + \beta X$$

The slope β is determined by the correlation ρ, and the standard deviations σ_X and σ_Y :

$$\beta = \rho\, \sigma_Y/\sigma_X$$

The correlation between Y and X is zero if and only if the slope is zero.

Also note that, when Y and X have a bivariate normal distribution, the conditional variance of Y, given X, is constant ie not a function of X:

$$Var(Y|X) = \sigma^2_{Y|X}$$

This is why, in the usual linear regression model $Y = \alpha + \beta X + \varepsilon$, the variance of the "error" term ε does not depend on X.

However, not all variables are linearly related. Suppose we have two random variables related by the equation $S = T^2$ where T is normally distributed with mean zero and variance 1. Then the linear correlation between S and T is zero, even though one is a function of the other.

Linear correlation is a measure of how close two random variables are to being *linearly* related. In fact, if we know that the linear correlation is +1 or -1, then there must be a deterministic linear relationship $Y = \alpha + \beta X$ between Y and X (and vice versa).

If Y and X are linearly related, and f and g are functions, the relationship between f(Y) and g(X) is not necessarily linear, so we should not expect the linear correlation between f(Y) and g(X) to be the same as between Y and X.

For a study of correlations between two lognormals see the paper " A common misconception about correlations".
click here