How to Use the Correlation and P.I.G. Spreadsheet

We assume standardized random variables X, Y with Gaussian distributions.

Use this spreadsheet to input any linear correlation R into any row of Column B.

The equivalent Percentage Information Gain (P.I.G.) will be output in Column G.

For example – input R = .387. Output will be 5.9%.

Why does a linear correlation correspond exactly to one Percentage Information Gain? Well…

Percentage Information Gain

= I(X;Y) / H(Y)

= H(Y) – H(Y|X) / H(Y)
[(the original entropy) – (the entropy of the model error)] / [the original entropy]

= H(Gaussian with standard deviation 1) – H(Gaussian with standard deviation $\sqrt{1-R^2}$) / H(Gaussian with standard deviation =1).

= (2.05 – (2.05 + (log2($\sqrt{1-R^2}$)) / (2.05)

= – (log2($\sqrt{1-R^2}$)/2.05)

The formula in Column G.