Hi Mark,

I just wanted to clarify your claim yesterday that Pearson's linear correlation coefficient (ρ) = the slope (β) estimated from a linear least squares fit of $y = \beta x + \alpha$ on the same data.

For two random variables x and y, the linear correlation coefficient is defined:

$$\rho = \frac{\left\langle \left(x_i - \langle x \rangle\right)\left(y_i - \langle y \rangle\right) \right\rangle}{\sqrt{\left\langle \left(x_i - \langle x \rangle\right)^2 \right\rangle \left\langle \left(y_i - \langle y \rangle\right)^2 \right\rangle}} = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}$$

The slope derived using linear least squares (minimising MSE etc..) with y regressed on x:

$$\beta = \frac{\left\langle \left(x_i - \langle x \rangle\right)\left(y_i - \langle y \rangle\right) \right\rangle}{\left\langle \left(x_i - \langle x \rangle\right)^2 \right\rangle} = \frac{\mathrm{cov}(x, y)}{\sigma_x^2}$$

$$\Rightarrow \boxed{\rho = \beta \left(\frac{\sigma_x}{\sigma_y} \right) \qquad (1)}$$

Two points are noteworthy:

1. Eqn (1) holds for any two random variables in general ($\sigma_y \neq 0$) and $\beta = 0 \Rightarrow \rho = 0$ naturally as you claim.
2. However, if we transform the x and y data to z-scores so that they have zero mean and unit variance, ie:

$x_i \rightarrow x_i' = \dfrac{x_i - \langle x \rangle}{\sigma_x}$ and $y_i \rightarrow y_i' = \dfrac{y_i - \langle y \rangle}{\sigma_y}$, then $(\sigma_{x'} / \sigma_{y'}) = 1$ and we get:

$$\boxed{\rho = \beta \qquad (2)}$$

when ρ and β are both computed from the new data x_i' and y_i'.

So it's important to note that when one has no knowledge of how two datasets are distributed, one cannot immediately claim that $\rho = \beta$. If the x, y data are related by a scale factor (eg. the price of diamonds versus the salaries of the people who mine them), then β subsumes this scale dependence and it's possible that $|\beta| > 1$. In this case, their standard deviations (or relative spreads) are needed to estimate ρ via eqn (1). With proper transformation of the data to z-scores, your claim is correct!

Cheers,
Frank