Pearson Correlation Coefficients and R-squared Explanation

We start with two random variables X and Y, now notice the covariance: $Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ has units so there is no interpretation. Therefore, we have to standardize both variables defined as such:

$$Z_X = \frac{X - \mu_X}{\sigma_X}, \quad Z_Y = \frac{Y - \mu_Y}{\sigma_Y} \tag{1}$$

Once we standardized the variables, the most natural way to measure their linear association is to take the dot product. To come up with the dot product, let us suppose we have n observations:

$$Z_X = (z_{x1}, z_{x2}, ..., z_{xn}), \quad Z_Y = (z_{y1}, z_{y2}, ..., z_{yn})$$

The dot product of the two vectors is:

$$\begin{aligned} Z_X \cdot Z_Y &= \sum_{i=1}^{n} z_{xi} z_{yi} \\ &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \\ &= \frac{1}{s_x s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Where $z_{xi} = \frac{x_i - \bar{x}}{s_x} \quad and \quad z_{yi} = \frac{y_i - \bar{y}}{s_y}$, (note we have switched to sample notation from population notation)

If we divide it by n:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} z_{xi}z_{yi} &= \frac{1}{s_x s_y}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= Sample\, estimate\, of\, \mathbb{E}\left[Z_X Z_Y\right]\end{aligned}$$

The motivation for taking the dot product illustrates an underlying geometric meaning, recall that for any two vectors u and v, the dot product is related as such where the cosine of the angle between the two vectors tell us how aligned they are:

$$cos\theta = \frac{u \cdot v}{\|u\| \, \|v\|} \tag{2}$$

In our case:

$$\begin{aligned}cos\theta &= \frac{Z_X \cdot Z_Y}{\|Z_X\| \, \|Z_Y\|} \\ cos\theta &= \frac{\sum_{i=1}^{n} z_{xi}z_{yi}}{\|Z_X\| \, \|Z_Y\|} \\ &= \frac{\sum_{i=1}^{n} z_{xi}z_{yi}}{\sqrt{\sum_{i=1}^{n} z_{xi}^2}\sqrt{\sum_{i=1}^{n} z_{yi}^2}}\end{aligned}$$

Now notice:

$$\begin{aligned}
z_{xi} &= \frac{x_i - \bar{x}}{s_x} \\
z_{xi}^2 &= \frac{(x_i - \bar{x})^2}{s_x^2} \\
\sum_{i=1}^{n} z_{xi}^2 &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s_x^2} \\
&= \frac{1}{s_x^2} \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

By definition of standard deviation (divide by n version):

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = n s_x^2$$

Therefore,

$$\sum_{i=1}^{n} z_{xi}^2 = \frac{1}{s_x^2} n s_x^2 = n, \quad \sqrt{\sum_{i=1}^{n} z_{xi}^2} = \sqrt{n}$$

Therefore,

$$\begin{aligned}
cos\theta &= \frac{\sum_{i=1}^{n} z_{xi} z_{yi}}{\sqrt{n}\sqrt{n}} \\
&= \frac{1}{n} \sum_{i=1}^{n} z_{xi} z_{yi}
\end{aligned}$$

We actually do not have to do all this calculation but just know that $cos\theta$ is related to the dot product.

Therefore now we define Pearson Correlation Coefficient as:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \quad written\, as\, r\, in\, tutorial\, notes \tag{3}$$

Because notice:

$$\begin{aligned}\rho &= \mathbb{E}\left[Z_X Z_Y\right] \\ &= \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}\end{aligned}$$

Which is conceptually just the dot product, and cosine of the angle between the standardized data vectors, telling us how aligned are X and Y

The numerator $Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ captures how X and Y behave together. If both X and Y are above their means, then the deviations are positive, which means the product is positive. If both X and Y are below their means, then the deviations are negative, but since a negative times a negative is positive, the product is positive. If one is above the mean and the other is below, the product is negative. This product tells us whether the two variables tend to move in the same direction (both increasing or both decreasing) or in opposite directions.

We then divide $Cov(X,Y)$ by $\sigma_X \sigma_Y$ so that the number tells us how strongly X and Y move together relative to how much they each normally vary. The size of covariance depends on how big the variables are, not just how related they are. Covariance tells us whether X and Y are moving in the same direction but leaves out the question of how strong is their relationship. That comparison requires dividing by their standard deviations.

Now we defined the R-squared value as:

$$R^2 = \rho^2 \tag{4}$$

Because we are only concerned with magnitude of the correlation between X and Y, not direction, which is captured also by the cosine of the angle.