Equivalence Pearson coefficient and OLS estimator

Context

When computing linear regression using the ordinary least squares (OLS) estimator between 2 variables, we are doing a correlation analysis, even if it's not explicited tough. However, it exists a very famous correlation measure: Pearson correlation coefficient.

If you have already use both of these tools, you might have already tell yourself that they it feel like linear regression and correlation coefficient describe a similar thing. And that is completely true. In fact, there is an equivalence between Pearson correlation coefficient and the estimator of the OLS method, and it's very simple:

If we assume the following regression model: $y = \beta_0 + \beta_1 x + \varepsilon$, we have:

$$\hat{\beta}_1 = cor(x, y) \times \frac{\sigma_y}{\sigma_x}$$

Example

Let's see with a concrete example:

```
x = rnorm(n=100)
y = 2*x + rnorm(100)
cor(x,y)
```

```
## [1] 0.877703
```

This is the correlation between our variables. Now let's compute the coefficient from the regression:

```
reg = lm(y ~ x)
reg$coefficients[2]
```

```
##        x
## 1.778093
```

And this is what happen when we multiply it by the standard deviations quotient:

```
cor(x,y) * sd(y)/sd(x)
```

```
## [1] 1.778093
```

Nice they seem the same thing: it's because they are. But where does it come from? Let's see it in more detail.

Proof

First, let's do the proof of the definition of the estimator that minimizes the sum squared error:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad and \quad r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x$$

$$Argmin_\beta \quad Loss = L(y, x, \beta) = \sum_1^n r_i^2 = \sum_1^n (y - \beta_0 - \beta_1 x)^2$$

$$\frac{\partial L}{\partial \beta_0} = -2\sum_1^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum y_i = \sum_1^n \hat{\beta}_0 + \sum_1^n \hat{\beta}_1 x_i = n\hat{\beta}_0 + \sum_1^n \hat{\beta}_1 x_i$$

$$\frac{1}{n}\sum_1^n y_i = \frac{1}{n} n\hat{\beta}_0 + \frac{1}{n}\sum \hat{\beta}_1 x_i$$

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

We use this last result for the following:

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial \sum_1^n (y - \beta_0 - \beta_1 x)^2}{\beta_1}$$

$$= \frac{\partial \sum_1^n (y_i - (\overline{y} - \beta_1 \overline{x}) - \beta_1 x_i)^2}{\beta_1} = 0$$

$$= \frac{\partial \sum_1^n [(y_i - \overline{y}) - \beta_1 (x_i - \overline{x})]^2}{\beta_1} = 0$$

$$-2\sum_1^n [(y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})]^2 = 0$$

$$-2\sum_1^n [(x_i - \overline{x})((y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x}))] = 0$$

$$\sum_1^n (x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta}_1 \sum_1^n (x_i - \overline{x})^2$$

$$\frac{1}{n}\sum_1^n (x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta}_1 \frac{1}{n}\sum_1^n (x_i - \overline{x})^2$$

$$cov(x, y) = \hat{\beta}_1 Var(x) <=> \hat{\beta}_1 = \frac{cov(x, y)}{Var(x)}$$

From this and with the fact that $Var(x) = \sigma_x^2$ and $cor(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$, we can re-write the latter:

$$\hat{\beta}_1 = \frac{cov(x, y)}{Var(x)} = \frac{cov(x, y)}{\sigma_x \sigma_x} = \frac{cov(x, y)}{\sigma_x \sigma_y}\frac{\sigma_y}{\sigma_x} = cor(x, y)\frac{\sigma_y}{\sigma_x}$$

I hope that this post has shed some light on the relationship between the ordinary least squares estimator and the Pearson correlation coefficient and why they are in fact the same thing.