Random Variables and Stochastic Process (AI5030)

Soumyajit Chatterjee
AI22MTECH02005

March 8, 2022

Question 55 Dec (2018)

Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3) \ldots (X_n, Y_n)$ be n independent observations from a distribution. Let r_p be the product moment correlation coefficient and r_s be the rank correlation coefficient computed based on n observations. Which of the following statements is correct.

1. $r_p \geqslant 0$ implies $r_s \geqslant 0$
2. $r_s \geqslant 0$ implies $r_p \geqslant 0$
3. $r_p = 1$ implies $r_s = 1$
4. $r_s = 1$ implies $r_p = 1$

Solution

Theoretical proof
The correlation coefficient for a bi-variate dataset with the independent vector as X and dependent vector as Y is given by

$$\rho = \frac{cov(X, Y)}{var(X) * var(Y)}$$

Where the covariance of X, Y in vector form is given as:

$$\text{cov(X,Y)} = \text{E}[(X - \overline{X})^T (Y - \overline{Y})]$$

Where the variance of X in vector form is given as:

$$\text{var(X)} = \text{E}[(X - \overline{X})^T (X - \overline{X})]$$

The product moment correlation coefficient r_p is also given as:

$$\rho = \frac{cov(X, Y)}{var(X) * var(Y)}$$

The assumption made while calculating r_p is that there exists some linear relation among the X and Y variables in the data.

Here, linear relation means a straight line relationship exists between X and Y.

The rank correlation coefficient r_s is given as:

$$\rho = \frac{cov(R(X), R(Y))}{var(R(X)) * var(R(Y))}$$

Where R( ) returns the rank/index of the sorted data as appearing in the dataset.
The assumption made while calculating r_s is that there exists some monotonic relation among the X and Y variables in the data.

Here, monotonic relation means some increasing relationship exists between X and Y. Example, Y can be increasing linearly with X, Y can be increasing parabolically with X, Y can be increasing exponentially with X.

Without loss of generality we can say that a linear function is a subset of monotonic functions.

Therefore, $r_p \geqslant 0$ means positive linear correlation which means as X increases, Y also increases linearly.

$r_s \geqslant 0$ means positive monotonic correlation which means as X increases, Y also increases monotonically.
If a function increases linearly it must be increasing monotonically as well as linear functions are a subset of monotonic functions.

Therefore, from the above logic option (1) and option (3) are correct i.e $r_p \geqslant 0$ implies $r_s \geqslant 0$ and $r_p = 1$ implies $r_s = 1$ which is also a case of option (1).

Proof From simulation

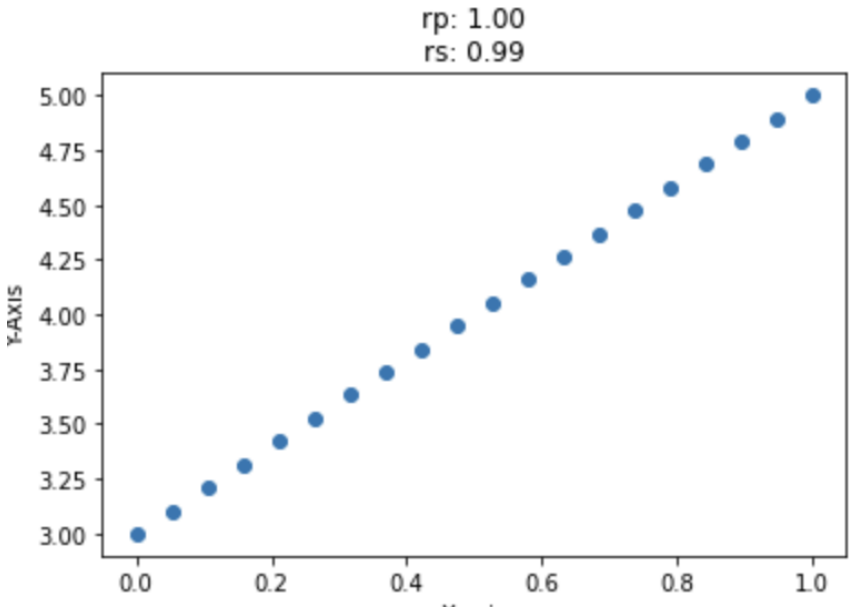

Figure 1: Linear Data

From the above plot we can see that since data is linear, r_p is positive and r_s is also positive and both of them are equal to 1 justifying option (1) and option(3) that for linear data r_p implies r_s.

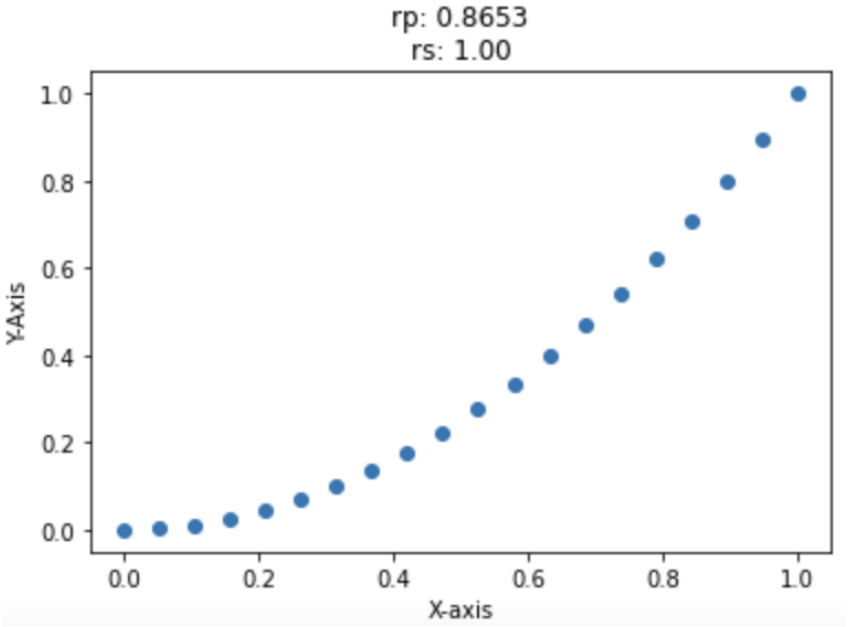

Figure 2: Exponential Data

From the above plot we can again see that if r_p is positive then r_s is also positive. But since data is not linear but simply monotonic, therefore r_p is not 1 but r_s is still 1 as r_s assumes any monotonic relation between the data points.

Therefore, $r_p = 1$ implies $r_s = 1$ but not the other way around as every linear function is monotonic but every monotonic function is not linear. Therefore this plot also justifies option (1) and option (3).