recipe_05

December 18, 2023

1 Anscombe's Quartet

The raw data has four series. The correlation coefficients are high. Visualization shows that a simple linear regression model is misleading.

1.1 Raw Data for the Series

The source file has four series, identified as `I`, `II`, `III`, and `IV`.

Each series has 11 (x, y) sample pairs.

1.2 Statistical Computations

Two properties compute correlation coefficient, r, and perform linear regression. The regression gives us two parameters for a line. - Slope, m. - Intercept, b.

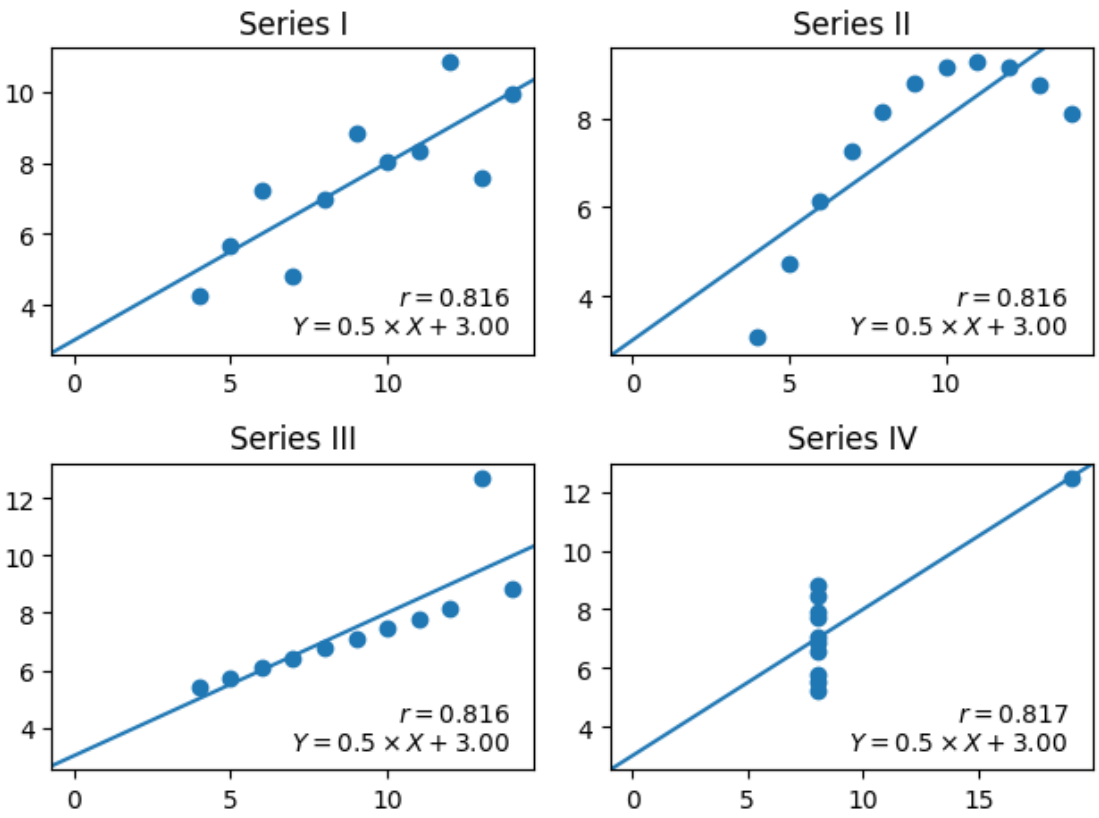
The equation for a line is $y = m \times x + b$

We can see that $r = 0.82$; this is a strong correlation.

This leads to a linear regression result with $y = 0.5 \times x + 3.0$ as the best fit for this collection of samples. Interestingly, this is true for all four series in spite of the dramatically distinct scatter plots.

1.3 Visualization

The following figures show the four series.


Series I
$r = 0.816$
$Y = 0.5 \times X + 3.00$
Series II
$r = 0.816$
$Y = 0.5 \times X + 3.00$
Series III
$r = 0.816$
$Y = 0.5 \times X + 3.00$
Series IV
$r = 0.817$
$Y = 0.5 \times X + 3.00$