Covariance and Correlation

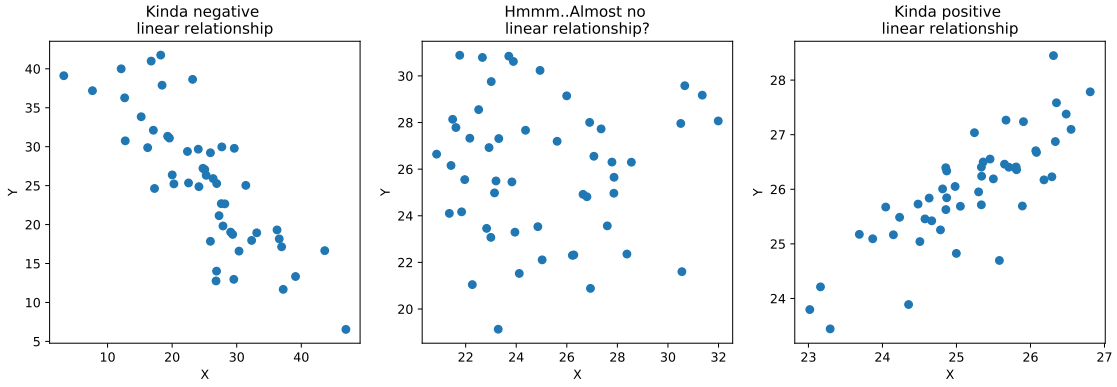
Parthiban Rajendran

October 31, 2018

1 Why

Earlier in regression, we said, by eyeballing, one could roughly conclude if a viable regression line possible that could be useful. But that of course, is not a rigorous approach to decide upon the **goodness** of relation between two variables. Note that for all below variation in X and Y, we could still draw a regression line, but it is obvious, for those **closer** to linear relationship between them positively or negatively will benefit from regression line than those who do not.

We need a rigorous reliable mathematical measure for linear relationship between X and Y



2 What

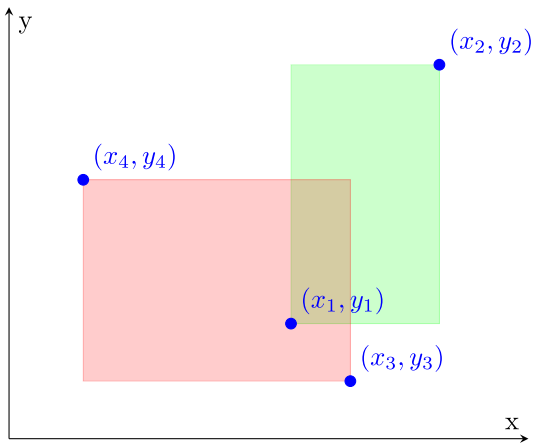
Relationship Definition

Let X and Y be the random variables involved, and each point representing a (x, y) pair value. What we want to see is, how is each point located with respect to every other point in the given sample set. Also we want to know if that is in a positive or negative way. Imagine a pair of points (x_1, y_1) and (x_2, y_2). Let x_1 and x_2 be in increasing order, then if $(y_2 > y_1)$ we could say, the pair is in a positive relationship. We could also sort $y_1, y_2, \cdots$ in increasing order, and then say if $x_2 > x_1$, then the pair is in a positive relationship. By positive we just mean, with increasing x the y increases. The negative relationship is defined simply the opposite of it, that is, with increasing x, the y decreases. Or with increasing y, the x decreases. Consequently, in terms of points we could say, given $y_1 < y_2$, if $x_1 > x_2$, then its a negative relationship. Summarizing we could stick to below convention, but one could try the alternate also.

Given $(x_1, y_1), (x_2, y_2)$ and y is in increasing order, i.e., $(y_1 < y_2)$,
if $(x_1 < x_2)$ or $(x_2 - x_1 > 0)$, this implies x has increased with y, a positive relationship
if $(x_1 > x_2)$ or $(x_1 - x_2 > 0)$, this implies x has decreased with y, a negative relationship

Visual Quantification via Colored Rectangles

Now that we have defined the relationship, next should think about quantification. After all, what we seek is a *measure*, a quantification of the relationship. How could we quantitatively differentiate the defined relation between pairs say, $[(x_1, y_1), (x_2, y_2)]$ and $[(x_3, y_3), (x_4, y_4)]$? This could be approached with geometry. Imagine drawing a rectangle based on $[(x_1, y_1), (x_2, y_2)]$, say R_{12} and $[(x_3, y_3), (x_4, y_4)]$, say R_{34} separately. Then one rectangle's area would be smaller or larger than the other, indicating a quantified measure of how farther apart the points are comparitively. Also, we could color the area to indicate if the involved pair that is used to construct the rectangle is in a positive or negative relationship. To construct a rectangle out of two points $[(x_1, y_1), (x_2, y_2)]$, we could just consider them as a two oppositing corners of the rectangle, and simply draw one whose sides are parallel to the axes. Let us color green for a positive relationship and red for a negative relationship. Such a visual quantification is illustrated below. Note that, a certain transparency is maintained for each rectangle, so the overlapping does not hide any information, but simply transparent to us.



$$y_1 < y_2 < y_3 < y_4$$
$$x_1 < x_2 \text{ so } (x_1, y_1) \text{ +ve with } (x_2, y_2)$$
$$x_3 > x_4 \text{ so } (x_3, y_3) \text{ -ve with } (x_4, y_4)$$

$$\text{Area of a rectangle, } R_{ij} = (x_i - x_j)(y_i - y_j) \tag{1}$$

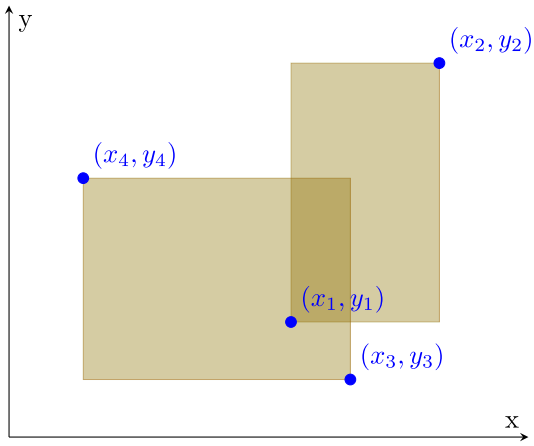
3 Area Distribution

Of course we have not drawn all possible combinations above for given set of points $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$ to establish first the basic idea, but that is what we would do for any given set of points:Plot all such relationship rectangles for every point with every other point in the given sample. We want to know for each and every point in given set, its relationship with every other point, *quantitatively*. However there is a problem.

> *If we try to plot for every possible pair of given set of data, there will be symmetrically distributed duplicity which not only introduces redundancy in the measure, but also neutrlizes our visualization*

That is, if (x_i, y_i) is a positive relationship with (x_j, y_j), it also means (x_j, y_j) is in negative relationship with x_i, y_i in other direction. Trying to take all possible rectangle will have this duality for all rectangles. For example, by iterated dual looping if at one iteration if $(x_i, y_i) = (1, 2), (x_j, y_j) = (3, 4)$, then down the

line, when j takes i value, we have, $(x_i, y_i) = (3, 4), (x_j, y_j = (1, 2))$. In terms of rectangle notation, for every R_{ij}, there is R_{ji} which is of equal value.



Duality Issue nullifying rectangles

Thus the flaw in the visualization already strongly suggests not to take all rectangles for the measure but may be, just half of it as representative of entire sample set. Below are the total number of rectangles for $N = 6$ pairs of sample sets. The blue shaded is symmetrical to yellow shaded. This is why the measure would be inherently doubled if all rectangles are taken into account. By nature, it is not needed. Think about it. Taking all possible rectangles, simply means, looking for a linear relationship in one direction and then again, in reverse, and deciding that the relationship is null. We should instead decide to take in to account only one direction,which means, only half of below rectangles would sufficely give a measure of relationship in one direction. Also note the diagonal rectangles have zero area, thus can be neglected too.

y_6	R_{61}	R_{62}	R_{63}	R_{64}	R_{65}	R_{66}
y_5	R_{51}	R_{52}	R_{53}	R_{54}	R_{55}	R_{56}
y_4	R_{41}	R_{42}	R_{43}	R_{44}	R_{45}	R_{46}
y_3	R_{31}	R_{32}	R_{33}	R_{34}	R_{35}	R_{36}
y_2	R_{21}	R_{22}	R_{23}	R_{24}	R_{25}	R_{26}
y_1	R_{11}	R_{12}	R_{13}	R_{14}	R_{15}	R_{16}
y / x	x_1	x_2	x_3	x_4	x_5	x_6

$N = 6$,
Total rectangles = N^2

Thus, we would just go with only either blue or yellow rectangles as illustrated above. Let us look closer at the product $(x_i - x_j)(y_i - y_j)$ for all rectangles. The no of rectangles in the half we are interested in

is given by $\frac{N(N-1)}{2}$. If $N = 6$, you could observe we have $\frac{(6)(5)}{2} = 15$ rectangles as our interest out of $N^2 = 6^2 = 36$ rectangles.

If we untangle the rectangle information systematically,we could come up with a summation to calculate the total value as below. Let us consider the *yellow* rectangles (you could try the blue ones)

- Let $i = 1$, then $R_{12} + R_{13} + R_{14} + R_{15} + R_{16} = \sum_{j=i+1}^{6} R_{1j}$
- Let $i = 2$, then $R_{23} + R_{24} + R_{25} + R_{26} = \sum_{j=i+1}^{6} R_{2j}$
- Let $i = 3$, then $R_{34} + R_{35} + R_{36} = \sum_{j=i+1}^{6} R_{3j}$
- Let $i = 4$, then $R_{45} + R_{46} = \sum_{j=i+1}^{6} R_{4j}$
- Let $i = 5$, then $R_{56} = \sum_{j=i+1}^{6} R_{5j}$

We could thus consilidate the total area of our interest as,

$$\text{Total Interested Area, TIA} = \sum_{i=1}^{5} \sum_{j=i+1}^{6} R_{ij}$$

When $i = 6$, $j = i + 1 = 7$, and there is no R_{67}, or $R_{67} = 0$, so we could rewrite slightly as,

$$\text{TIA } = \sum_{i=1}^{6} \sum_{j=i+1}^{6} R_{ij}$$

Using 1, and generalizing to N,

$$\text{TIA} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j) \tag{2}$$

Alternate approach: We instead could have taken all area, and then simply divided by 2. Here, the derivation is straight forward. For $N = 6$, there are $N^2 = 36$ rectangles possible. And as indexed in last diagram, the total area would be,

$$\text{Total Area} = \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij}$$

Using 1 and taking the half as that is our interested area, we get,

$$\text{TIA} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j) \tag{3}$$

Both 2 and 3 are equivalent, but 2 gives a better intuition, what we are after. Let us take a closer look next at the rectangular area distribution.