Covariance and Correlation

Parthiban Rajendran
parthi292929@gmail.com

November 14, 2018

Contents

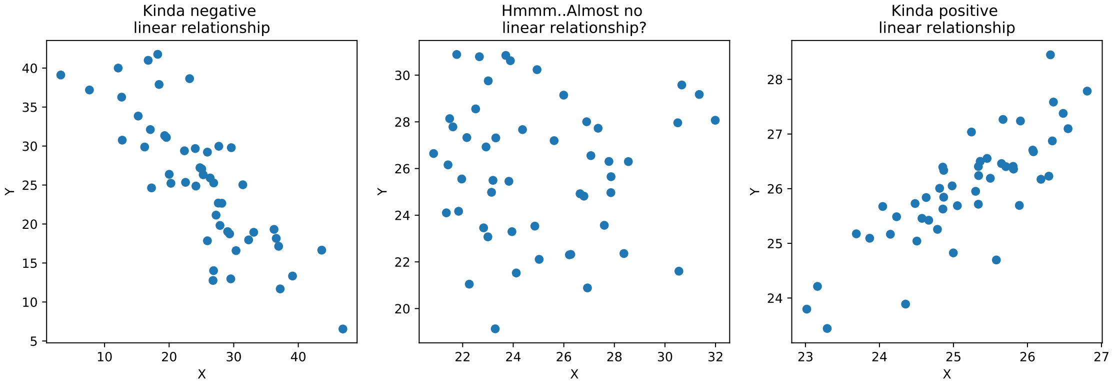
Chapter 1

Covariance

1.1 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



1.2 What

Relationship Definition

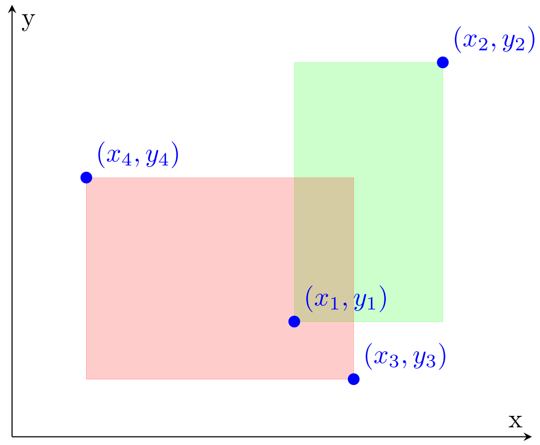
Let X and Y be the random variables involved, and each point representing a (x, y) pair value. What we want to see is, how is each point located with respect to every other point in the given sample set. Also we want to know if that is in a positive or negative way. Imagine a pair of points (x_1, y_1) 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.



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

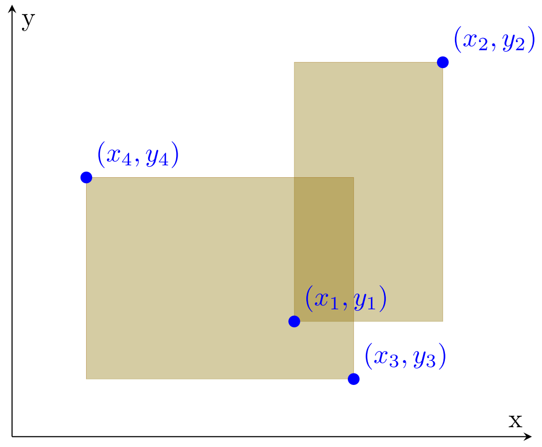
$$\text{Area of a rectangle, } R_{ij} = (x_i - x_j)(y_i - y_j) \tag{1.1}$$

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.

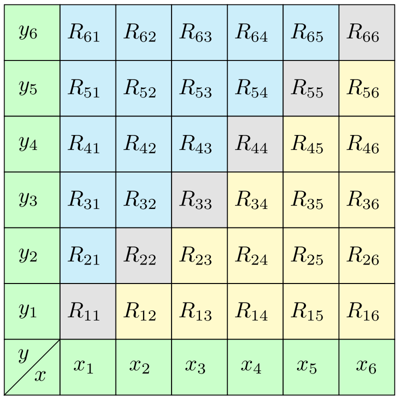
> *If we try to plot for every possible pair of given set of data, there will be symmetrically distributed duplicity which not only introduces redundancy in the measure, but also neutrlizes our visualization*

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

$$\begin{aligned} N &= 6, \\ \text{Total rectangles} &= N^2 \end{aligned}$$

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 $\dfrac{N(N-1)}{2}$. If $N = 6$, you could observe we have $\dfrac{(6)(5)}{2} = 15$ 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 1.1, and generalizing to N,

$$\text{TIA} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j) \tag{1.2}$$

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

$$\text{TIA} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j) \tag{1.3}$$

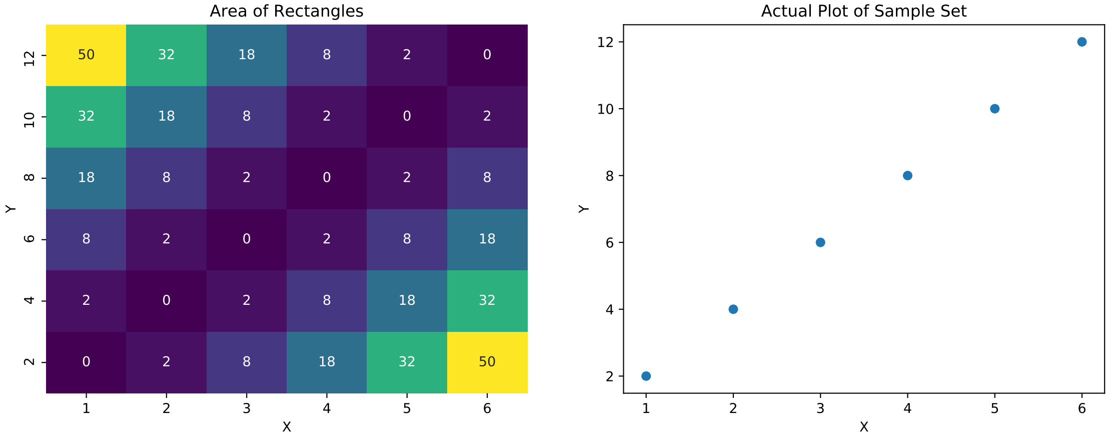
Both 1.2 and 1.3 are equivalent, but 1.2 gives a better intuition, what we are after. Let us take a closer look next at the rectangular area distribution.

Case 1 : Perfectly positively linearly related dataset (whew!)

Suppose we have such a case as below. you could note, this is of line $y = 2x$

- X = 1,2,3,4,5,6
- Y = 2,4,6,8,10,12

Then, every possible rectangle for each pair of $[(x_i, y_i), (x_j, y_j)]$ is tabulated below. This illustrates the redundancy better. Note the repetitive values symmetrically spread from the diagonal lines. The color gradient gives a better perspective of the spread. The actual plot of the sample set is given on the right side.



Using 1.2 or 1.3, TIA for given sample set, turns out to be 210

```
In[33]: X , Y= [1,2,3,4,5,6],[2,4,6,8,10,12]
        N = len(X)

        def get_TIA(X,Y):
            N = len(X)
            comb_l, area = sorted(zip(X,Y), key=lambda x: x[1]), 0  #sorting w.r.t Y
            for i in range(0,N):  # equivalent for i = 1 to N because, range is 0 to N-1
                for j in range(i+1,N):
                    X1, Y1, X2, Y2 = comb_l[i][0], comb_l[i][1], comb_l[j][0], comb_l[j][1]
                    d1, d2 = X2 - X1, Y2 - Y1
                    area += d1*d2
            return area

        print(get_TIA(X,Y))
```

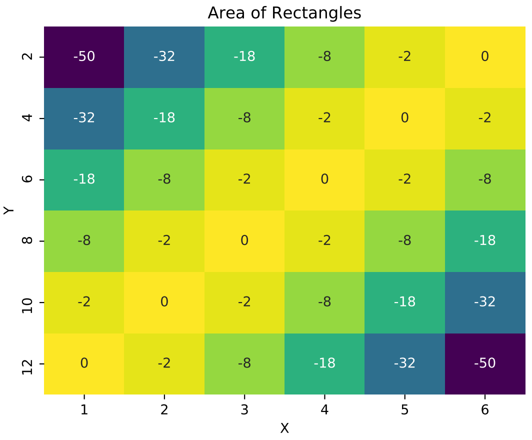
```
210
```

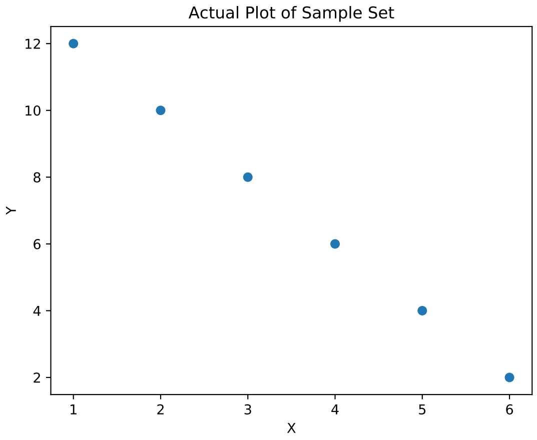
Case 2 : Perfectly negatively linearly related dataset

Suppose we have such a case as below. you could note, this is of line $y = 14 - 2x$

- X = 1,2,3,4,5,6
- Y = 12,10,8,6,4,2

For this, let us check the rectangles' area.





We see something interesting here. If you might have thought, some way the rectangular area also represented actual plot could notice it here that the area plot on left hand side is still similarly symmetrical as before, even though the plot is perfectly negatively related as shown on RHS. This is because, that was its definition in first place. The rectangular area plot on LHW just gives a measure of the spread of relationship, while the plot on RHS represents the actual location. Also note, that again, due to symmetry, we have duplicate values, thus suggesting to halve the measure. And the values are negative. This is good, now that could help to indicate our sample sets are negatively linearity related. Let us check out the TIA.

```
In[35]: X , Y= [1,2,3,4,5,6],[12,10,8,6,4,2]
        print(get_TIA(X,Y))
```

```
-210
```

Its negative. We are already getting somewhere! Let us consider another case, where there is no linear relationship.

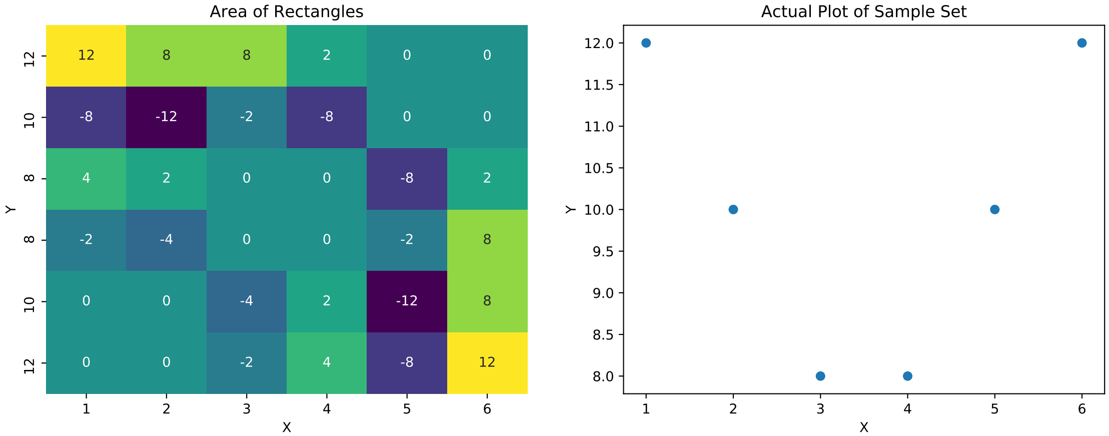
Case 3 : Dataset with no linear relationship

Suppose we have such a case as below.

- X = 1,2,3,4,5,6

- Y = 12,10,8,8,10,12

The respective rectangle area and plots are as below.



Again, irrespective of actual plot, the area graph on LHS, is still symmetrical if you look carefully, assuring, no matter what, the measure is available in doubled quantity across all possible rectangles, so good golly gosh, we chose half of the rectangles. Note the RHS plot, there is clearly not a possibility of a best fit linearity between X and Y, and this should reflect in our measure. Let us calculate the TIA.

```
In[37]: X , Y = [1,2,3,4,5,6],[12,10,8,8,10,12]
        print(get_TIA(X,Y))
```

```
0
```

Understandly it is 0. The no linear relationship in a literal sense has been transformed to a number via our TIA. Recall,

- for a perfectly positively linearly related dataset, we got +210
- for a perfectly negatively linearly related dataset, we got -210
- for a perfectly not linearly related dataset, we got 0

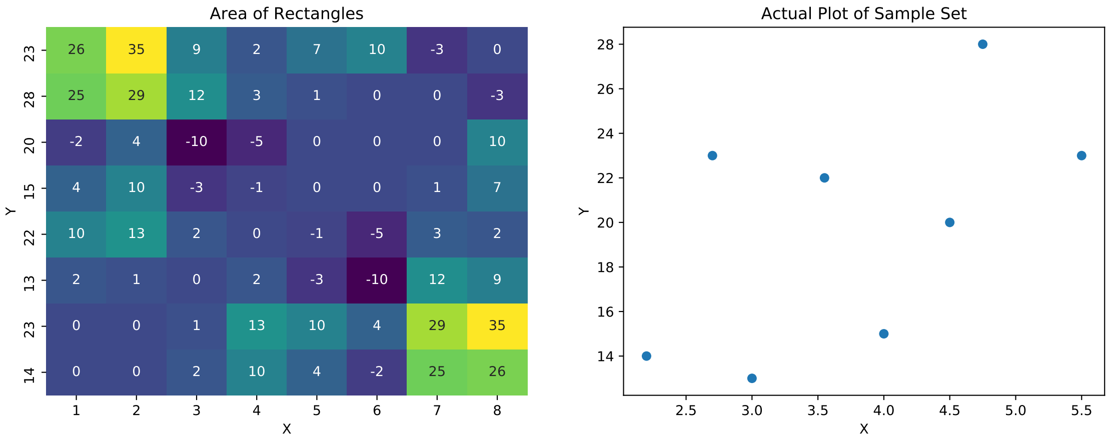
Thus our TIA is already proving to be a good measure. Note that, if we had taken all rectangles and getting 420,-420,0 instead, it would be an unnecessarily doubled stretch, giving a doubled sense of actual linearly underneath. By halving the area, that is via TIA, we have taken the linearity sense in a kind of *same* scale of what it is.

Case 4: A practical realistic dataset

Suppose we have such a case as below.

- X = 2.2, 2.7, 3, 3.55, 4, 4.5, 4.75, 5.5
- Y = 14, 23, 13, 22, 15, 20, 28 , 23

The respective rectangle area and plots are as below.



Y \ X	1	2	3	4	5	6	7	8
23	26	35	9	2	7	10	-3	0
28	25	29	12	3	1	0	0	-3
20	-2	4	-10	-5	0	0	0	10
15	4	10	-3	-1	0	0	1	7
22	10	13	2	0	-1	-5	3	2
13	2	1	0	2	-3	-10	12	9
23	0	0	1	13	10	4	29	35
14	0	0	2	10	4	-2	25	26

You see, even for a realistic sample set which has some linearity associated in either direction (positive or negative), the LHS area diagram has a symmetry as usual. This would always be the case, thus we are right in taking the half no of rectangles, no matter what the linearity is. Proceeding to TIA, we get it as

```
In[39]: X , Y = [2.2, 2.7, 3, 3.55, 4, 4.5, 4.75, 5.5],[ 14, 23, 13, 22, 15, 20, 28 , 23]
        print(get_TIA(X,Y))
```

```
184.39999999999995
```

```
The tikzmagic extension is already loaded. To reload it, use:
  %reload_ext tikzmagic
```

1.4 Visualization

Now that we have seen TIA is already doing a good job on giving us a measure of the linearity, we shall come to the core of this section. We have not yet visualized the totality of the rectangles. We could have done this earlier, but I wanted to instill a strong sense of what rectangles are we dealing with and why they are whole representative of the dataset though we have taken only half of all possible rectangles. We initially decided how do we color the rectangles, based on positive or negative relationship as a convention, and then looked in detail, what are the rectangles to be plotted. Let us consider the sample sets as below. Recall these were the same sample sets we saw in the beginning of this section. Note the TIA is already calculated indicating us the kind of relationship. As per TIA from Figure 1.2, Plot 1 is highly negatively correlated, Plot 2 is somewhat

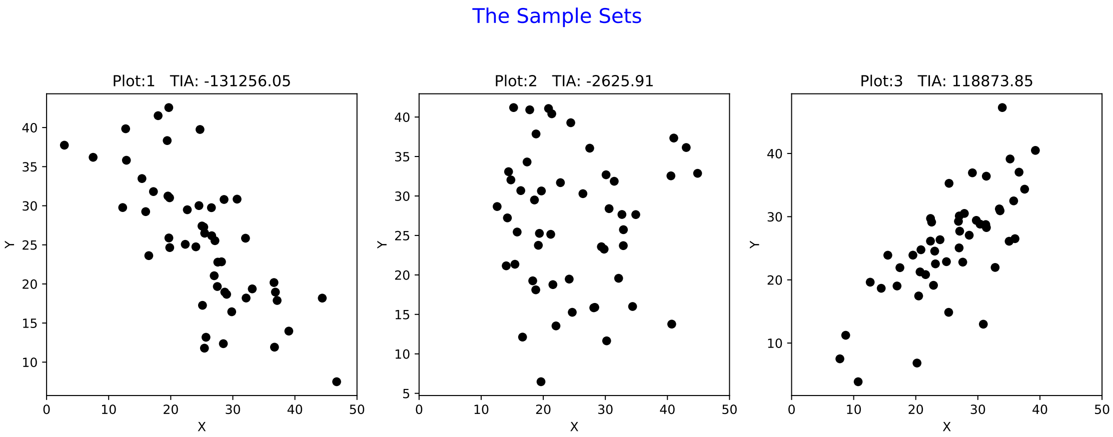


Figure 1: The Sample Sets

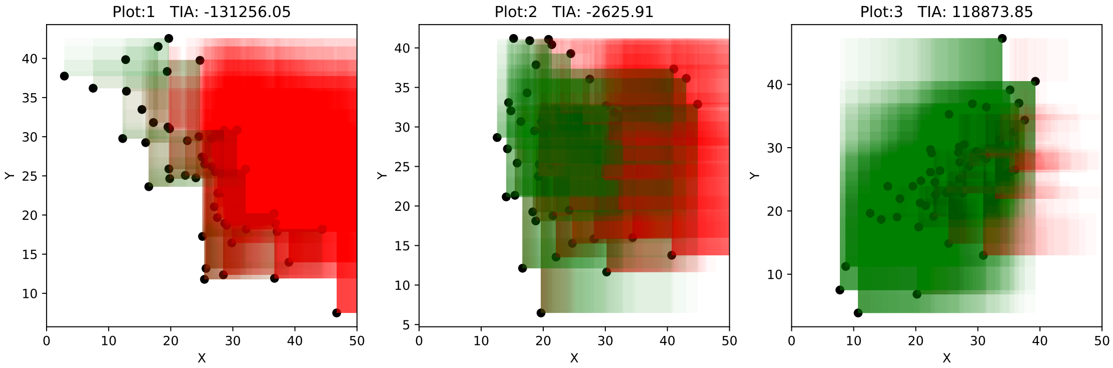


Figure 2: The Visualization of Covariance

negative, and Plot 3 is positively correlated. Though visually Plots 1 and 3 look like not having much difference in their *slope* or *rate*, our TIA gives a wide difference in value. This is because, TIA between sample sets are not comparable (we will solve that soon in correlation, but remember this problem). That is, given a sample set, say Plot 1, having -148859 is one of infinite no of possibilities among that sample set, with perfectly linear positive, negative and 0 TIA as one of those. Simiarly for sample set in Plot 2 and so on. Below are the sample sample plots with colored rectangles laid over them. Remember, if N is the size of sample set, or no of (x, y) pairs, then the number of rectangles we have drawn is $N(N-1)/2$. And as we already saw, only because of this limited rectangles, we get the output as below without neutralization issues.

I think, Figure 1.2 speaks for itself :) Plot 1,which has highly negative linear relationship among its sample sets, has more red rectangles than green. Plot 2, which is very less linearity in any direction, shows an almost equal mix of red and green, of course the accurate measure is reflected in its TIA though. Plot 3, which has a positive linear relationship, obviously has lot more green. Figure 1.3 gives total area of red and green separately, giving us better glimpse of the *net* relationship underneath. The TIA is just the difference between the total green area and red area.

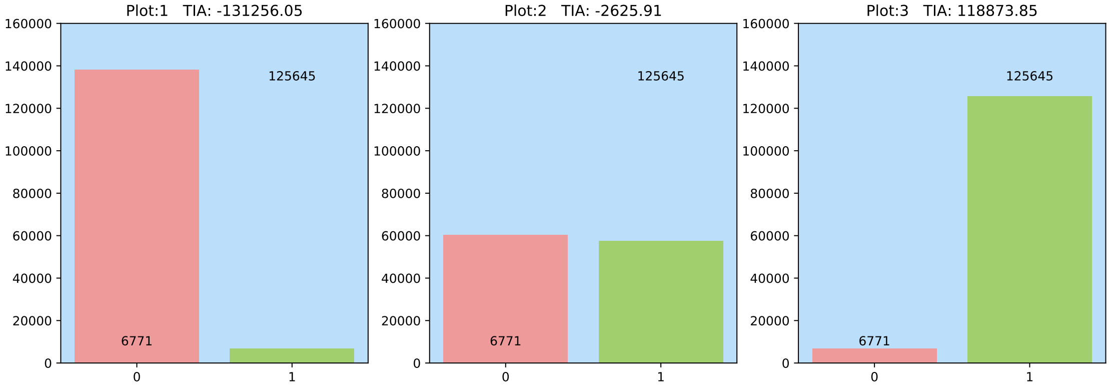


Figure 3: The separated total area
Green indicates Positive

1.5 Expected value of TIA

For any given sample set, we are typically interested not in the total of the sample set, but most probable or best representative candidate of that sample set. In our case, our sample set of TIA, is not individual pairs (x_i, y_i), but a function of them, a product $(x_i - x_j)(y_i - y_j)$. That is, using 1.2 if,

$$h(X, Y) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)$$

then, we are interested in $E[h(X, Y)]$
As per expectation formula,

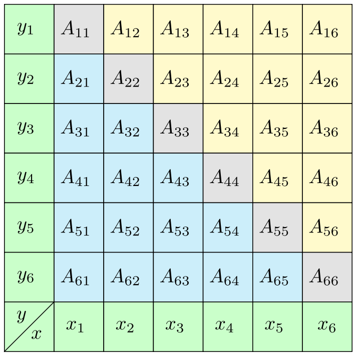
$$E[h(X, Y)] = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)p(x_i, y_i) \tag{1.4}$$

> Note, we are not interested in expected value of *number of rectangles* or *red colored rectangles* etc. The area of rectangles carry the measure and each rectangle might have different area. We are thus interested in the *expected value* of the area, given the *total* interested area.

Expectation needs a *joint probability mass function* $p(X, Y)$ associated with $h(X, Y)$. Recall the rectangle graph for $N = 6$ and replace with area A_{ij} (could also call as product, P_{ij} but just to avoid notational confusion with probability let us stick with area).

Assuming each *area* has equal probability, given the number of area, each A_{ij} will have a probability of $\frac{1}{N^2}$ as there are N^2 area components possible. Thus, 1.4 becomes,

$$E[h(X, Y)] = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)p(x_i, y_i) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j) \tag{1.5}$$



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

$$N \;=\; 6,$$
$$\text{Total rectangles} = N^2$$
$$\text{TIA} = \sum_{i=1}^{N}\sum_{j=i+1}^{N} A_{ij}$$

Figure 4: The Number of Area Components

Ladies and Gentlemen. That $E[h(X,Y)]$ is called **Covariance** of X and Y , shortly called Cov($\mathbf{X},\mathbf{Y}$). Also note, the alternative form we saw earlier in equation 1.3, could also be used to derive covariance as below.

$$E[h(X,Y)] = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(x_i - x_j)(y_i - y_j)p(x_i,y_i) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j) \tag{1.6}$$

Covariance of discrete X and Y with p(X Y) uniform

Given X and Y are discrete variables of sample size N, and $p(X,Y) = \dfrac{1}{N^2}$,

$$\text{Cov}(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(x_i - x_j)(y_i - y_j) \tag{1.7}$$

$$\text{Cov}(X,Y) = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)(y_i - y_j) \tag{1.8}$$

1.6 Standard Formula

What we have seen so far, is a deformed form of covariance which numerically gave us the same results as a standard formula. It is mathematically possible to show that,

$$\text{Cov}(X,Y) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(x_i - x_j)(y_i - y_j)p(x_i,y_i) = \sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})p(x_i,y_i) \tag{1.9}$$

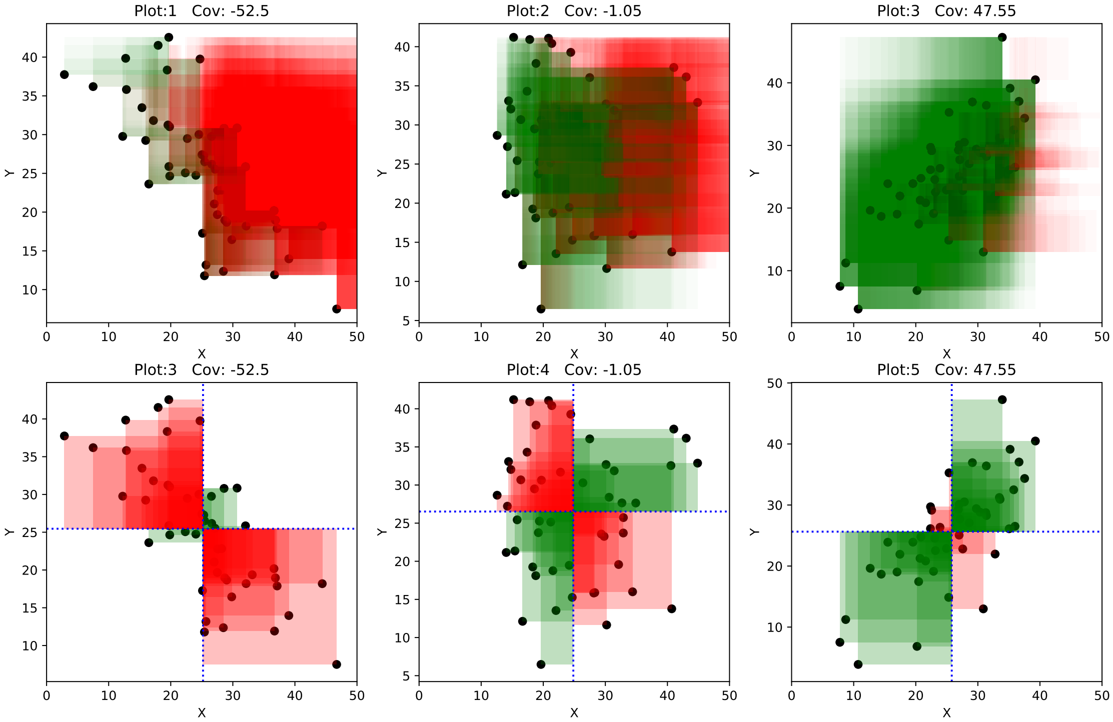


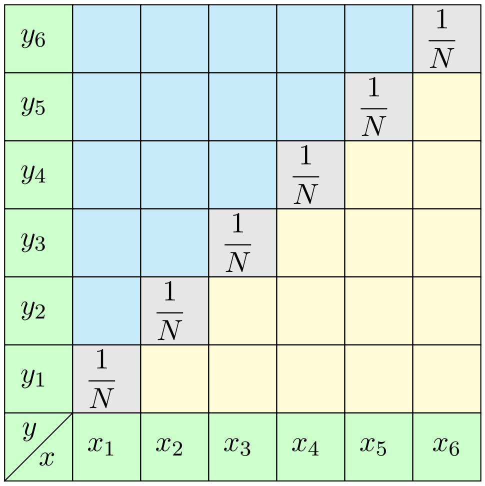
Figure 5: The Visualization of deformed and standard formula for Covariance

The derivation is proven by Zhang et al. [5]. At the time of this writing, the doubts in the derivation is not yet cleared, if and once it is done, this section should be enriched with a proper derivation. Till then, this is a discontinuity in our understanding. The visualization of standard formula is slightly different because it involves mean, so all rectangles have one corner at mean position $(\overline{x}, \overline{y})$. The visualization is shown in figure 1.5. The top 3 rows from our deformed formula and bottom 3 using standard formula. One could observe, the rectangles in plots 3,4,and 5 are centered around the mean (shown in dotted lines), thus giving a better visual perception of the measure (no of red or green rectangles, which is more). We did not start with this visualization only because, there was no intuition to introduce mean in the equation out of no where.

1.7 Generalization

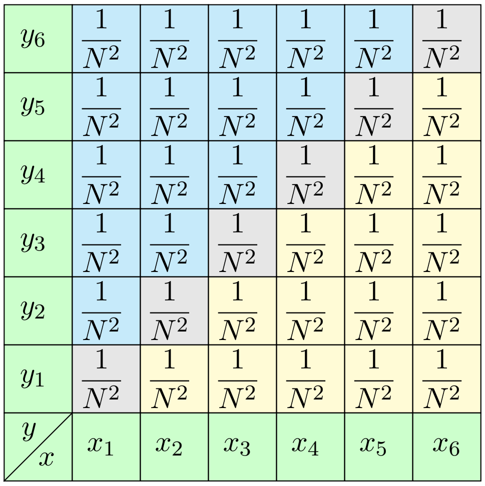
So far we have seen Covariance for discrete X, Y random variables. This could easily be transferred to continuous variables as well. However before generalization of the formula, we need to generalize the way the sample set is provided as well.

Suppose the sample set is given as $(X, Y) = (x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_N, y_N)$ then, if we say equi probable, then $p(X, Y)$ could be simply tabulated in different ways depending on the function $h(X, Y)$ that is, if we take the deformed or standard formula. This is illustrated in figure 1.6. This was simply because, of the way we indexed the sample points. In Plot A, we do not have a (x_2, y_1), because we just numbered as $(x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_N, y_N)$, and it worked because standard formula needed only one time indexing via i. But in Plot B, we had double indexing via i, j, this



y_6						$\frac{1}{N}$
y_5					$\frac{1}{N}$	
y_4				$\frac{1}{N}$		
y_3			$\frac{1}{N}$			
y_2		$\frac{1}{N}$				
y_1	$\frac{1}{N}$					
y / x	x_1	x_2	x_3	x_4	x_5	x_6

Plot A
$p(X,Y)$ for standard formula
$h(X,Y) = (X - \overline{X})(Y - \overline{Y})$



y_6	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_5	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_4	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_3	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_2	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_1	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y / x	x_1	x_2	x_3	x_4	x_5	x_6

Plot B
$p(X,Y)$ for deformed formula
$h(X_i, Y_i, X_j, Y_j) =$
$(X_i - X_j)(Y_i - Y_j)$

$p(X,Y)$ depending on $h(X,Y)$

is why the probability at each *cell* also became $1/N^2$. Most often we do not use the deformed formula and stick to standard formula. Further, often the given probability density function (if given), would be something like this.

		y		
$\rho(x,y)$		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

Here our indexing style has to differ. Now we have $(x_1, x_2) = (100, 250)$ and $(y_1, y_2, y_3) = (0, 100, 200)$. If we line up these sample pairs, we get

$$(x_1, y_1), (x_1, y_2), (x_1, y_3), (x_2, y_1), (x_2, y_2), (x_2, y_3)$$

Thus even with standard formula due to data being in a different format, we would need to use double summation in order to vary i and j to different limits separately. Thus naturally our standard formula would become

$$\text{Cov}(X,Y) = \sum_{i=1}^{2} \sum_{j=1}^{3} (x_i - \overline{x})(y_j - \overline{y}) p(x_i, y_i)$$

Generalizing the standard formula, and also extending to continuous X and Y, we could say,

Generalized Standard Covariance Formula

The **covariance** between two rv's X and Y is

$$\text{Cov}(X,Y) = E[(X-\mu_x)(Y-\mu_Y)]$$

$$= \begin{cases} \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) & \text{X,Y discrete} \\ \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,y)dxdy & \text{X,Y continuous} \end{cases} \tag{1.10}$$

Depending on samples are from population or we deal with entire population, either $\overline{x}$ or μ_X could be used respectively.

1.8 Example

We have already explained the concept with an example, so here will see a different approach.

Suppose joint and marginal pmf's for X = automobile policy deductible amount and Y = homeowner policy deductible amount are as below. Find the covariance.

			y	
$\rho(x,y)$		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

x	100	250
$\rho_X(x)$	.5	.5

y	0	100	200
$\rho_Y(y)$	.25	.25	.5

This example was taken from Devore [1] Since we need the means in the equation, let us calculate them first.

$$\mu_x = \sum_{i=1}^{2} x_i p_X(x_i) = 100(0.5)+250(0.5) = 175\mu_y = \sum_{i=1}^{2} y_i p_Y(y_i) = 0(0.25)+100(0.25)+200(0.5) = 125$$

Coming to Covariance,

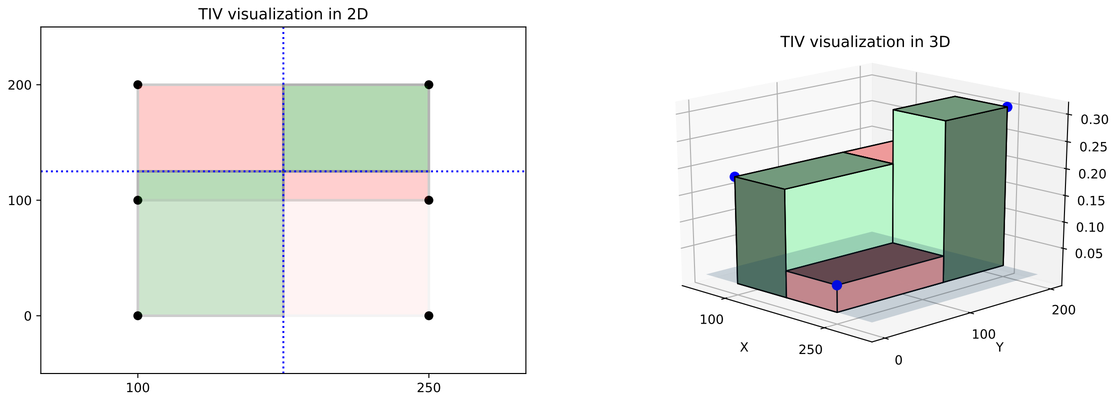


Figure 6: The Visualization of standard formula in 2D and 3D

$$\text{Cov}(X,Y) = \sum_{i=1}^{2}\sum_{j=1}^{3}(x_i-\mu_x)(y_j-\mu_y)p(x_i,y_j)$$

$$= (x_1-175)(y_1-125)p(x_1,y_1) + (x_1-175)(y_2-125)p(x_1,y_2) + (x_1-175)(y_3-125)p(x_1,y_3)$$
$$+(x_2-175)(y_1-125)p(x_2,y_1) + (x_2-175)(y_2-125)p(x_2,y_2) + (x_2-175)(y_3-125)p(x_2,y_3)$$

$$= (100-175)(0-125)p(100,0) + (100-175)(100-125)p(100,100) + (100-175)(200-125)p(100,200)$$
$$+(250-175)(0-125)p(250,0) + (250-175)(100-125)p(250,100) + (250-175)(200-125)p(250,200)$$

$$= (100-175)(0-125)0.20 + (100-175)(100-125)0.10 + (100-175)(200-125)0.20$$
$$+(250-175)(0-125)0.05 + (250-175)(100-125)0.15 + (250-175)(200-125)0.30$$

$$= 1875$$

What just happpened? How come we took all possible pairs of (x, y) given in joing pmf as *samples*? Earlier, when we visualized TIA for random samples, we assumed that $h(X, Y)$ had equal probability for all of its values, thus resulting in a constant probability for entire summation. So it was enough if we look at it from the sky or top or whatever. If the probability density in the summation is a variable, then just by looking at 2D, we are missing the *contribution* of pmf to the summation. Now that we have varying pmf for different pairs of x, y, we need to account for that, because pairs having higher probability will attract more samples than those that would not, thus potentially forming a relationship between X and Y. This is evident the moment we visualize in 3D as shown in figure 1.7. In 3D, it is evident now, the green has more volume, than red, so we could expect higher samples in these region than the red, thus suggesting in fact a *positive* correlation. Thus, yeah it is no more just a TIA ,but **total interested volume, TIV**. Also, a pmf resembles all possible values of (x, y), so could imagine, sample set of all possible values in any multiples (1 occurance per pair, or 10 occurance per pair, etc).

Generalized Standard Covariance Visualization

The better generalized visualization of standard covariance formula is in volume, if underlying joint probability density function is not a constant.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \sum_x \sum_y (x_i-\overline{x})(y_i-\overline{y})p(x_i,y_i) \\ &= (x_1-\overline{x})(y_1-\overline{y})p(x_1,y_1)+\cdots+(x_i-\overline{x})(y_i-\overline{y})p(x_i,y_i)+\cdots \\ &= V_{11}+\cdots+V_{ij}+\cdots\end{aligned} \tag{1.11}$$

Chapter 2

Correlation

2.1 Why

Covariance has some painful disadvantages. There is no standard scale with which we could compare and say, the number obtained is high correlation. When we measure, say a distance of 10m, we do not just have the measure 10, we also *understand the size* of it because we have a standard scale for 1m. This allows us to compare with another distance, say 15m, and accurately understand the difference between them. This type of **standardization** or normalization is missing in our Covariance value.

Further, it is highly unit dependent as we are just multiplying two RVs of different units (the 3rd factor probability we multiply with, anyway is unitless). This means, if units change, our measure also could drastically change. Imagine the last example. If X and Y, the deductibles were in cents, then they just scale by 100 times in the summation. Note what this leads to.

$$\begin{aligned}
&\mathrm{Cov}(X,Y) = \sum_x \sum_y (100x - 17500)(100y - 12500)p(x,y) \\
&= (100)(100)\sum_x \sum_y (x - 175)(y - 125)p(x,y) \\
&= 10000(1875) \\
&= 18750000 \quad \text{cents}^2
\end{aligned}$$

Apart from a very high value, note the ugly units tag sticking with it. Though a covariance could give us a measure, this is not as useful as a unit like meters. Ideally, we would wish, our measure is units independent. Summarizing,

Covariance's main disadvantages

- Critically dependent on units of random variables being compared
- Not comparable with other covariance values

2.2 What

The idea to tackle the issue is by, well as said, *standardization or normalization with something*, thereby making it a ratio, due to which the units cancel out between numerator and denominator.

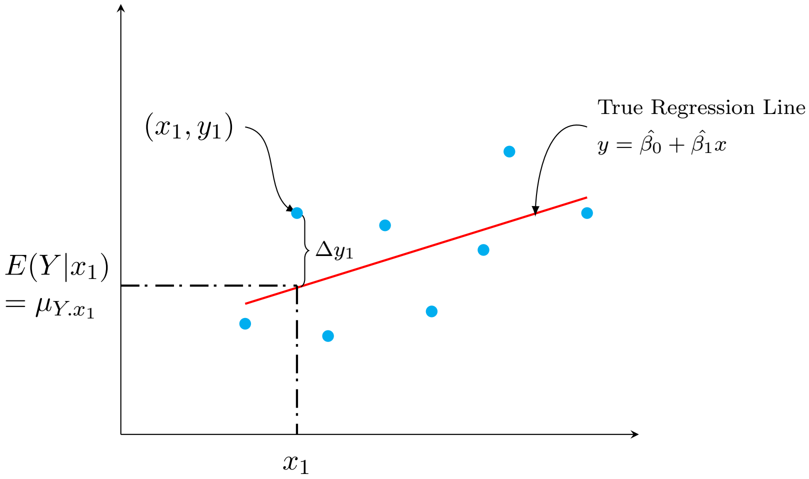


Figure 2.1: Recalling the regression line

This already suggests we need two quantities of same units of X and Y in the denominator of Covariance. Let us recall the equation of simple linear regression model between two Random variables (figure 2.1).

The regression line is given by

$$\begin{aligned} E(Y|x) &= \beta_0 + \beta_1 x \\ \hat{Y}|x &= \hat{\beta_0} + \hat{\beta_1} x \\ \text{where} \quad \hat{\beta_1} &= \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\ \hat{\beta_0} &= \overline{y} - \hat{\beta_1}\overline{x} \end{aligned} \tag{2.1}$$

What would it mean, when the slope $\hat{\beta_1}$ is 0 for this regression line?

$$\begin{aligned} \hat{\beta_1} &= 0 \\ \implies \hat{Y}|x &= \hat{\beta_0} = \overline{y} \end{aligned}$$

This is simply an horizontal line drawn parallel to x axis, cutting at $y = \overline{y}$. So, if such is the case, that for given sampe, $\hat{\beta_1}$ is 0, we could already say, their covariance is 0, because for any x, y remains constant at $\overline{y}$. This is illustrated in Figure 2.2.

Note in case of regression line, we took a variable X and evaluated the relationship of another variable Y via $E(Y|x)$. Thus naturally the reversed case is also possible that is $E(X|y)$. This is simply achieved by reversing the variables in regression line equation 2.1

$$\begin{aligned} E(X|y) &= \beta_2 + \beta_3 y \\ \hat{X}|y &= \hat{\beta_2} + \hat{\beta_3} y \\ \text{where} \quad \hat{\beta_3} &= \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (y_i - \overline{y})^2} \\ \hat{\beta_2} &= \overline{x} - \hat{\beta_3}\overline{y} \end{aligned} \tag{2.2}$$

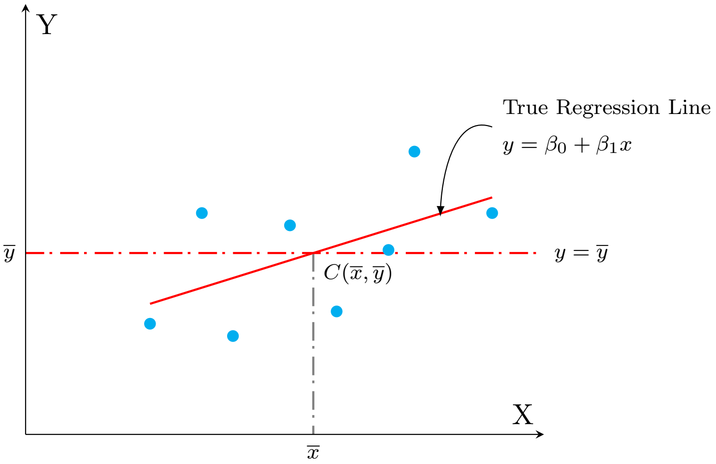


Figure 2.2: When the slope is zero..

Again, when $\beta_3 = 0$, that is slope of regression line $E(X|y)$ is 0, we get,

$$\hat{\beta}_3 = 0$$
$$\implies \hat{X}|y = \hat{\beta}_2 = \overline{x}$$

Figure 2.3 illustrates plotting of both lines, along with zero correlation lines.

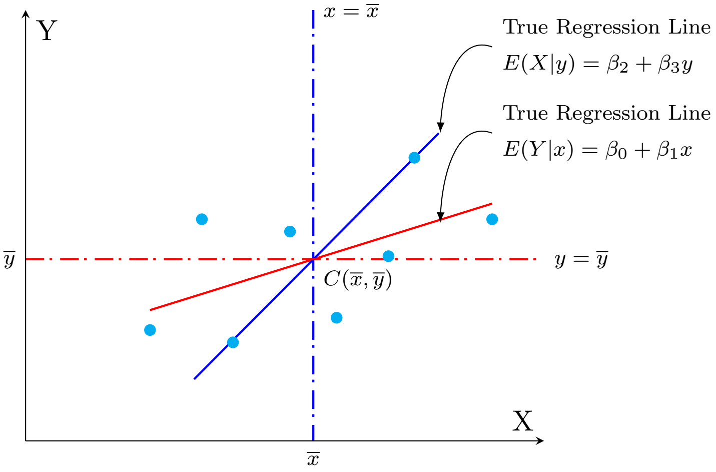


Figure 2.3: Two possible regression lines $E(Y|x), E(X|y)$

Summarizing, in current case of regression, we have,

- $E(Y|x)$ gives Y variation which is not same as variation indicated by $E(X|x)$
- $y = \overline{y}$ indicates zero variation of Y for any x, and $x = \overline{x}$, vice versa.

What we need is a single unified quantitative measure for reducing the disadvantages of Covariance. Note that we are dealing with samples, so our formula for *unbiased* sample covariance and variance, as referenced in Zaiontz [4], would be

$$\mathrm{cov}(X,Y) = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{x})(y_i - \overline{y})$$
$$\mathrm{var}(X) = s_x^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{x})^2$$
$$\mathrm{var}(Y) = s_y^2 = \frac{1}{N-1}\sum_{i}^{N}(y_i - \overline{y})^2$$

Using them in the slopes, we get,

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} = \frac{\frac{1}{N-1}\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{N-1}\sum_i (x_i - \overline{x})^2}$$
$$= \frac{\mathrm{cov}(X,Y)}{s_x^2}$$

Similary for $\hat{\beta}_3$. Summarizing, now we have, slopes in terms of sample covariance and variances,

$$\hat{\beta}_1 = \frac{\mathrm{cov}(X,Y)}{s_x^2} \quad , \quad \hat{\beta}_3 = \frac{\mathrm{cov}(X,Y)}{s_y^2} \tag{2.3}$$

Thus,

$$\hat{Y}|x = \hat{\beta}_0 + \frac{\mathrm{cov}(X,Y)}{s_x^2}x$$
$$\hat{X}|y = \hat{\beta}_2 + \frac{\mathrm{cov}(X,Y)}{s_y^2}y$$

Now, covariance is symmetric. X is as covariant with Y as Y is with X. Check the formula again.

$$\mathrm{cov}(X,Y) = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{N-1}\sum_{i}^{N}(y_i - \overline{y})(x_i - \overline{x}) = \mathrm{cov}(Y,X)$$

However, as we saw, this cannot be said for $\hat{Y}|x$ and $\hat{X}|y$. But imagine below form for a moment.

$$\hat{Y}|x = 0 + \frac{\mathrm{cov}(X,Y)}{1}x$$
$$\hat{X}|y = 0 + \frac{\mathrm{cov}(X,Y)}{1}y$$

If we some how magically make the y-intercept of $\hat{Y}|x$, and x-intercept of $\hat{X}|y$ go away, and make the variance 1, we could have a symmetry effect for both $\hat{Y}|x$ and $\hat{X}|y$. This could be done by *standardizing* the sample set. Recall during Z transformation, we did the same. By shifting the sample set or distribution to its mean, and scaling by the standard deviation, we essentially achieve a standard distribution which could be comparable to any other standardized distribution (Recall Z scores). Such a standardized distribution will have 0 mean and variance as 1.

Lemma

For a population described by RV, $X(\mu, \sigma^2)$

$$\begin{aligned}
Z &= \frac{X-\mu}{\sigma} \\
E(Z) &= E\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\big(E(X)-\mu\big) = \frac{1}{\sigma}(\mu-\mu) = 0 \\
\text{Var}(Z) &= \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \text{Var}\left(\frac{X}{\sigma}-\frac{\mu}{\sigma}\right) = \text{Var}\left(\frac{X}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1
\end{aligned}$$

Standardizing our sample set

Applying the same principles to our sample set, if we transform as follows,

$$X_s = \frac{X-\overline{x}}{s_X} \quad , \quad Y_s = \frac{Y-\overline{y}}{s_Y}$$

where s_X, s_Y are the standard deviation of X and Y respectively, then, we have new samples set (X_s, Y_s), where

$$\begin{aligned}
\overline{x_s} = \overline{y_s} &= 0 \\
s_{X_s} = s_{Y_s} &= 1
\end{aligned}$$

The new standardized set gives rise to new regression lines as follows.

$$\begin{aligned}
\hat{Y_s}|x_s &= \hat{\beta_{0s}} + \frac{\text{cov}(X_s, Y_s)}{s^2_{X_s}} x_s \\
\hat{X_s}|y_s &= \hat{\beta_{2s}} + \frac{\text{cov}(X_s, Y_s)}{s^2_{Y_s}} y_s
\end{aligned}$$

Using equations 2.1, and 2.2 we get,

$$\begin{aligned}
\hat{\beta_{0s}} &= \overline{x_s} - \hat{\beta_{1s}}\overline{y_s} = 0 - \hat{\beta_{1s}}(0) = 0 \\
\hat{\beta_{2s}} &= \overline{y_s} - \hat{\beta_{3s}}\overline{x_s} = 0 - \hat{\beta_{3s}}(0) = 0
\end{aligned}$$

Using that, and since $s_{X_s} = s_{Y_s} = 1$, we finally get new regression lines as,

$$\begin{aligned}
\hat{Y_s}|x_s &= \text{cov}(X_s, Y_s)x_s \\
\hat{X_s}|y_s &= \text{cov}(X_s, Y_s)y_s
\end{aligned}$$

Figure 2.4 illustrates the resultant regression lines. One could notice both these lines are symmetric because they both have same slope with respect to their independent axis.

The new *standardized* sample covariance $\text{cov}(X_s, Y_s)$ has very useful properties we have been longing so far.

- $\text{cov}(X_s, Y_s)$ would be now unitless and would vary between ±1 as we would observe shortly
- the covariance is now made symmetric, that is X_s is as covariant with Y_s as Y_s is with X_s
- this does not mean, the new regression lines are same. They just have same slope meaning they are *symmetric*

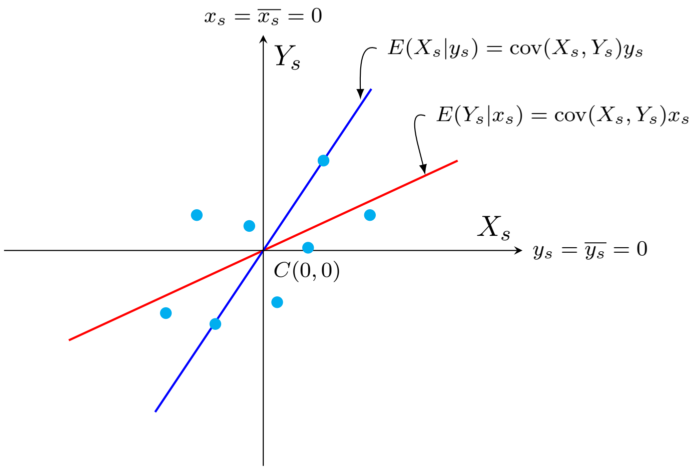

$x_s = \overline{x_s} = 0$
Y_s
$E(X_s|y_s) = \text{cov}(X_s, Y_s)y_s$
$E(Y_s|x_s) = \text{cov}(X_s, Y_s)x_s$
X_s
$y_s = \overline{y_s} = 0$
$C(0,0)$

Figure 2.4: Two standardized regression lines $E(Y_s|x_s), E(X_s|y_s)$

All the above points would become evident, once we observe a detailed example.

Covariance of Standardized Sample Sets

By standardizing the sample set, we are able to achieve interesting *symmetric* regression lines of same slope

$$\begin{aligned} \hat{Y_s}|x_s &= \text{cov}(X_s, Y_s)x_s \\ \hat{X_s}|y_s &= \text{cov}(X_s, Y_s)y_s \end{aligned} \tag{2.4}$$

where $\text{cov}(X_s, Y_s)$ is unitless and varies between ±1

2.3 Examples

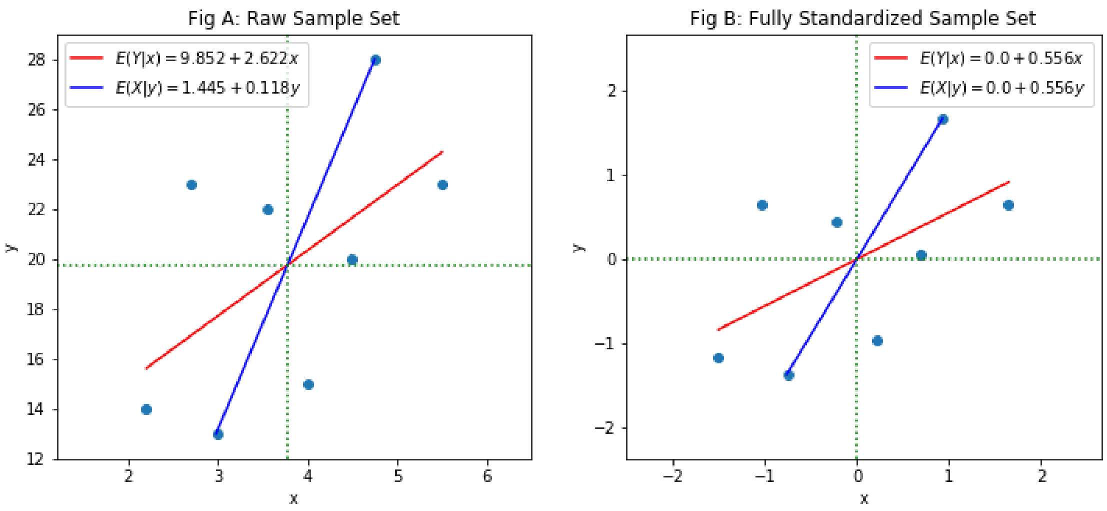
Example 1: A single simple sample set

Assume below is the given sample set. Let us plot both the direct simple regression line and standardized one to note the differences.

X	Y
2.2	14
2.7	23
3	13
3.55	22
4	15
4.5	20
4.75	28
5.5	23

```
In[17]: x_i = [2.2, 2.7, 3, 3.55, 4, 4.5, 4.75, 5.5]   # a sample set
         y_i = [14, 23, 13, 22, 15, 20, 28, 23]

         fig, axr = plt.subplots(1,2, figsize=(12,5))
         plot_regs(x_i, y_i, axr[0], std=False, label='Fig A: Raw Sample Set')
         plot_regs(x_i, y_i, axr[1], std=True, std_full=True, label='Fig B: Fully Standardized
         Sample Set')
         plt.show()
```



Note the regression line equations in both figures. In Figure B, as expected, both lines get the same slope which is *standardized covariance* $Cov(X_s, Y_s)$. Note the value of the common slope. It is positive and less than 1, this tells both sample sets are related linearly to an extent.

Example 2: Wikipedia Sample set

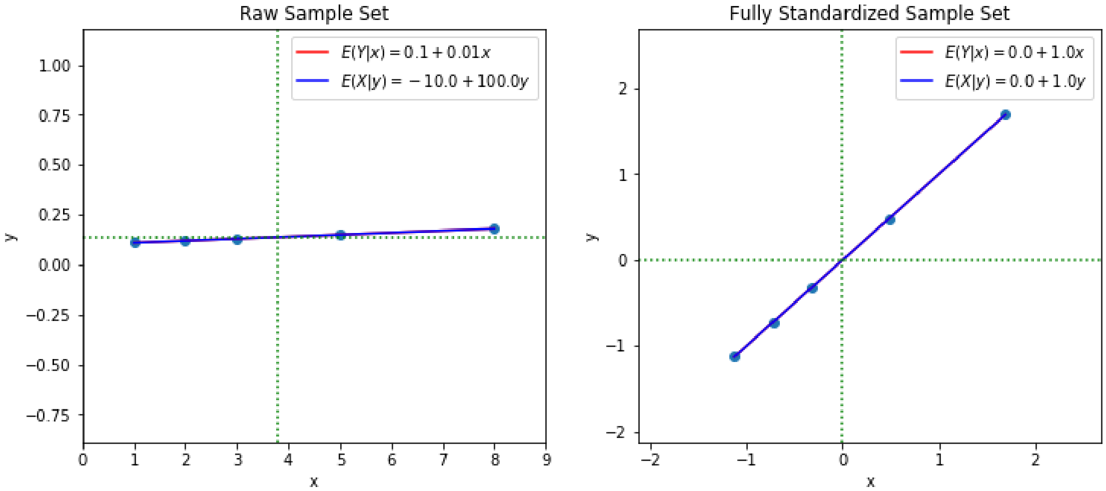
Let us try a perfectly covarying example. This is taken from Wikipedia's Pearson Correlation Coefficient article [1]

X	Y
1	0.11
2	0.12
3	0.13
5	0.15
8	0.18

```
In[18]: x_i = [1,2,3,5,8]    # a sample set
        y_i = [0.11,0.12,0.13,0.15,0.18]

        fig, axr = plt.subplots(1,2, figsize=(12,5))
        plot_regs(x_i, y_i, axr[0], std=False, label='Raw Sample Set')
        plot_regs(x_i, y_i, axr[1], std=True, std_full=True, label='Fully Standardized Sample
        Set')
        plt.show()
```

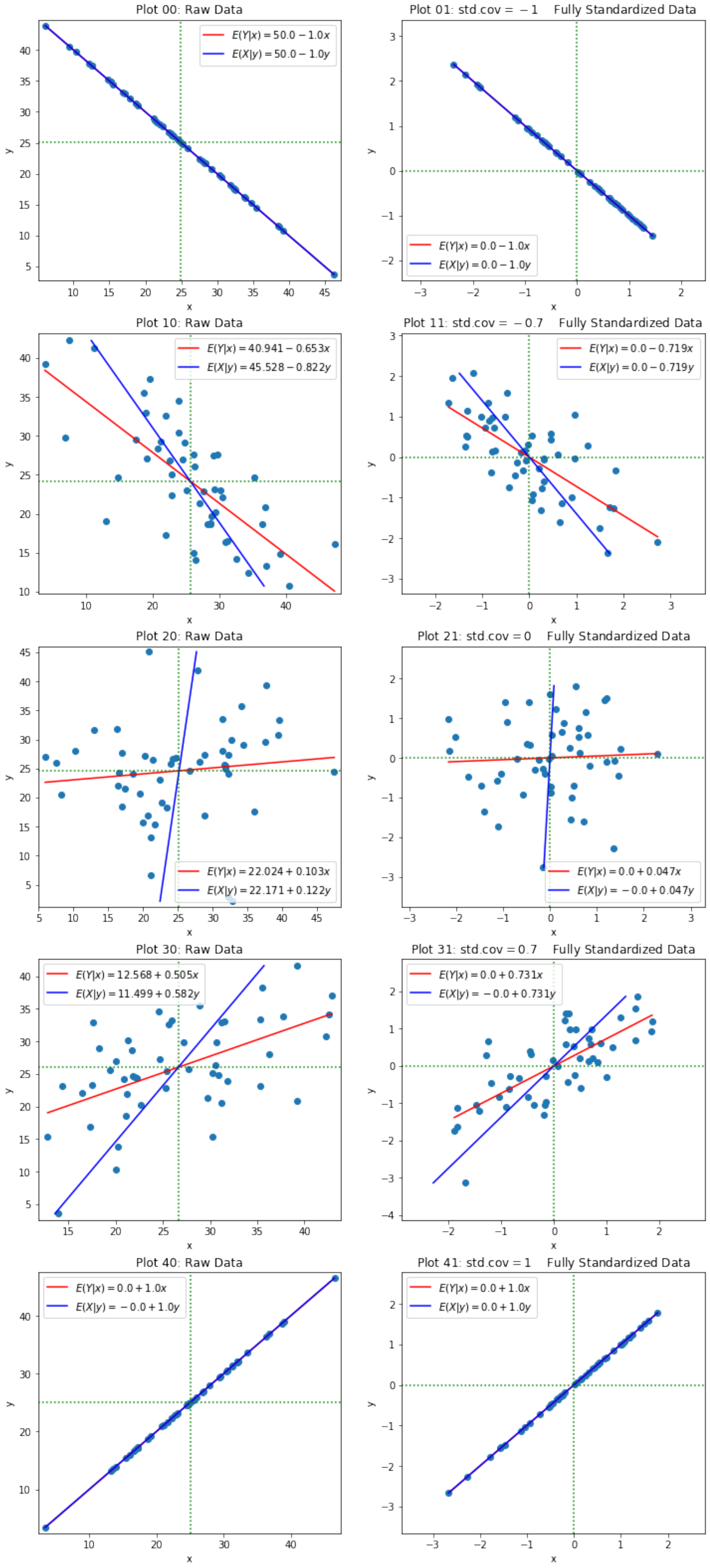
[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient



Aha! When the dataset is perfectly linearly related, we get the *standardized covariance* slope as 1. Ain't we getting somewhere?

Example 3: With different linear relationships

To test the different values of standardized covariance, we shall generate different datasets, that has perfect linearity in both directions (positive and negative), and also some what in the middle, including no linearity.


Plot 00: Raw Data
$E(Y|x) = 50.0 - 1.0x$
$E(X|y) = 50.0 - 1.0y$
x
y
Plot 01: std.cov = −1 Fully Standardized Data
$E(Y|x) = 0.0 - 1.0x$
$E(X|y) = 0.0 - 1.0y$
x
y
Plot 10: Raw Data
$E(Y|x) = 40.941 - 0.653x$
$E(X|y) = 45.528 - 0.822y$
x
y
Plot 11: std.cov = −0.7 Fully Standardized Data
$E(Y|x) = 0.0 - 0.719x$
$E(X|y) = 0.0 - 0.719y$
x
y
Plot 20: Raw Data
$E(Y|x) = 22.024 + 0.103x$
$E(X|y) = 22.171 + 0.122y$
x
y
Plot 21: std.cov = 0 Fully Standardized Data
$E(Y|x) = 0.0 + 0.047x$
$E(X|y) = -0.0 + 0.047y$
x
y
Plot 30: Raw Data
$E(Y|x) = 12.568 + 0.505x$
$E(X|y) = 11.499 + 0.582y$
x
y
Plot 31: std.cov = 0.7 Fully Standardized Data
$E(Y|x) = 0.0 + 0.731x$
$E(X|y) = -0.0 + 0.731y$
x
y
Plot 40: Raw Data
$E(Y|x) = 0.0 + 1.0x$
$E(X|y) = -0.0 + 1.0y$
x
y
Plot 41: std.cov = 1 Fully Standardized Data
$E(Y|x) = 0.0 + 1.0x$
$E(X|y) = 0.0 + 1.0y$
x
y

Note carefully.

- When the given dataset is perfectly negatively linearly related (Plot 00,01), $\text{cov}(X_s, Y_s) = -1$
- When the given dataset is somewhat negatively linearly related (Plot 10,11), $-1 < \text{cov}(X_s, Y_s) < 0$
- When the given dataset is totally not linearly related (Plot 20,21), $\text{cov}(X_s, Y_s) = 0$
- When the given dataset is somewhat positively linearly related (Plot 30,31), $0 < \text{cov}(X_s, Y_s) < 1$
- When the given dataset is perfectly positively linearly related (Plot 40,41), $\text{cov}(X_s, Y_s) = 1$

Thus, not only that our standardized covariance got rid of units, but also retains value between ±1, perfectly reflective of the linear relationship in the dataset. Thus we observe empirically via examples the range of standardized covariance.

2.4 Formalization of Sample and Population Correlation

The *standardized covariance* with its unique characteristic is thus called **Pearson's Correlation Coefficient**, **r** as it was formalized by Pearson. It is not required to standardize the sample set everytime, and calculate the standardized covariance as slope of the resultant regression line. We could calculate directly from the given sample set as below.

$$r = \text{cov}(X_s, Y_s) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{is} - \overline{x_s})(y_{is} - \overline{y_s})$$

Since standardized,

$$\overline{x_s} = \overline{y_s} = 0$$

$$x_{is} = \frac{x_i - \overline{x}}{s_X} \quad , \quad y_{is} = \frac{y_i - \overline{y}}{s_Y}$$

$$\begin{aligned}\therefore r = \frac{1}{N-1}\sum_{i=1}^{N}(x_{is})(y_{is}) &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right) \\ &= \frac{1}{s_X s_Y}\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \frac{\text{cov}(X,Y)}{s_X s_Y}\end{aligned}$$

Thus the *sample correlation coefficient* **r** of a given sample set (X, Y) is given by

$$r = \frac{\text{cov}(X,Y)}{s_X s_Y}$$

By analogy, a *population correlation coefficient* could also be derived. If (X, Y) are two discrete RVs, with $X = x_1, x_2, \cdots, x_N$, and $Y = y_1, y_2, \cdots, y_M$, and if $p(X, Y), p(X), p(Y)$ are their joint and marginal *pmf*s respectively, then a population correlation coefficient ρ could be defined as,

$$\rho = \frac{\sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(X,Y)}{\sqrt{\sum_x (x-\mu_X)^2 p(X) \sum_y (y-\mu_Y)^2 p(Y)}} \tag{2.5}$$

where, $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ are respective population parameters of X and Y. Recalling Covariance and Variance formula for population as below,

$$\begin{aligned}
\text{Cov}(X,Y) = \sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(X,Y) \\
\sigma_X^2 = \sum_x (x-\mu_X)^2 p(X) \\
\sigma_Y^2 = \sum_y (y-\mu_Y)^2 p(Y)
\end{aligned}$$

and using that, one could rewrite ρ as

$$\rho = \frac{\sum_x \sum_y (x-\mu_X)(y-\mu_Y)p(X,Y)}{\sqrt{\sum_x (x-\mu_X)^2 p(X) \sum_y (y-\mu_Y)^2 p(Y)}} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.6}$$

Sample and Population Correlation

The *sample correlation coefficient*, r of any given sample set (X,Y) is given by

$$r = \frac{\text{cov}(X,Y)}{s_X s_Y} \tag{2.7}$$

The *population correlation coefficient*, ρ of any given discrete RVs (X,Y) is given by

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2.8}$$

Similar ρ applicable to continuous RVs also, with integration suitably placed in place of summation.

2.5 Cosine Similarity

Interestingly correlation factor could be visualized to an extent in vector form or at least provides us easier computational method of calculation via matrices. Suppose there is a sample set (X,Y) of size 3. That is, if $(X,Y) = \{(x_1,y_1),(x_2,y_2),(x_3,y_3)\}$, we could represent them in a 3D vector form as below

$$\begin{aligned}
\vec{x} = x_1\hat{i} + x_2\hat{j} + x_3\hat{k} \\
\vec{y} = y_1\hat{i} + y_2\hat{j} + y_3\hat{k}
\end{aligned}$$

In simpler matrix notation,

$$\vec{x} = [x_1, x_2, x_3]$$
$$\vec{y} = [y_1, y_2, y_3]^T$$

Using law of cosines, the angle θ between vectors $\vec{x}, \vec{y}$ can be calculated as

$$\cos\theta = \frac{\vec{x} \bullet \vec{y}}{\|x\|\|y\|}$$

where

$$\vec{x} \bullet \vec{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \bullet \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 = \sum_i^3 x_iy_i$$

and

$$\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{\sum_i x_i^2}$$
$$\|y\| = \sqrt{y_1^2 + y_2^2 + y_3^2} = \sqrt{\sum_i y_i^2}$$

Readers are strongly advised to go through appendix 3.1 where the concept is explained in detail and also concluded that the above relation is applicable to any higher dimensional vector. Thus, recalling equation 3.9 from appendix, if the sample set size is N, then we could represent in matrix form and extend the cosine relationship as follows.

Let

$$\vec{x} = [x_1, x_2, x_3, \cdots, x_N]$$
$$\vec{y} = [y_1, y_2, y_3, \cdots, y_N]^T$$

then,

$$\vec{x} \bullet \vec{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \bullet \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = x_1y_1 + x_2y_2 + \cdots + x_Ny_N = \sum_i^N x_iy_i$$

and

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2} = \sqrt{\sum_i^N x_i^2}$$
$$\|y\| = \sqrt{y_1^2 + y_2^2 + \cdots + y_N^2} = \sqrt{\sum_i^N y_i^2}$$

so,

$$\cos\theta = \frac{\vec{x} \bullet \vec{y}}{\|x\|\|y\|} = \frac{\sum_i^N x_iy_i}{\sqrt{\sum_i^N x_i^2}\sqrt{\sum_i^N y_i^2}}$$

If we subtract the mean of the RVs, from each of the elements as below, setting up **centered** vectors,

$$\vec{x_c} = [x_1 - \overline{x}, x_2 - \overline{x}, x_3 - \overline{x}, \cdots, x_N - \overline{x}]$$
$$\vec{y_c} = [y_1 - \overline{y}, y_2 - \overline{y}, y_3 - \overline{y}, \cdots, y_N - \overline{y}]^T$$

this similarly leads to

$$\cos\theta = \frac{\vec{x_c} \bullet \vec{y_c}}{\|x_c\|\|y_c\|} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2}\sqrt{\sum_i^N (y_i - \overline{y})^2}}$$

which is same as *sample correlation coefficient*, r. Note that, the value of cosine ranges between ± 1. So when both vectors are in same direction, the θ is 0, thus $\cos\theta = 1$, maximum value indicating perfect linearity. Similarly when both vectors are in opposite direction, $\theta = 180°$, implying $\cos\theta$ = -1. When the vectors are perpendicular to each other, $\theta = 90°$ implying $\cos\theta = 0$, thus zero correlation.

For those, who find it difficult to comprehend higher dimensional vector, remember that in any higher dimensional vector, the angle between the resultant two vectors is always on a plane (2D), thus the law of cosine still applies. This is also explained in appendix 3.1

Cosine Similarity

The *sample correlation coefficient*, r of any given sample set (X, Y) can also be expressed in vector matrix form, giving a cosine relationship as

$$r = \cos\theta = \frac{\vec{x_c} \bullet \vec{y_c}}{\|x_c\|\|y_c\|} = \frac{\text{cov}(X, Y)}{s_X s_Y} \tag{2.9}$$

where, $\vec{x_c}$ and $\vec{y_c}$ indicate *centered* dataset

Chapter 3

Appendix

3.1 Dot Product

3.1.1 Angle between two 2D unit vectors

Suppose we have two unit vectors $\hat{u}, \hat{v}$ on a plane as shown in figure 3.1. We are interested in finding the angle between them θ which gives a measure of how much apart the vectors are. That is, quite a low angle, could say, both vectors are kind of in similar direction, and around 180 could mean, they are kind of in opposite direction and so on. We said unit vectors, but any vector to be called as a unit vector, it should satisfy the property that their *magnitude* is 1.

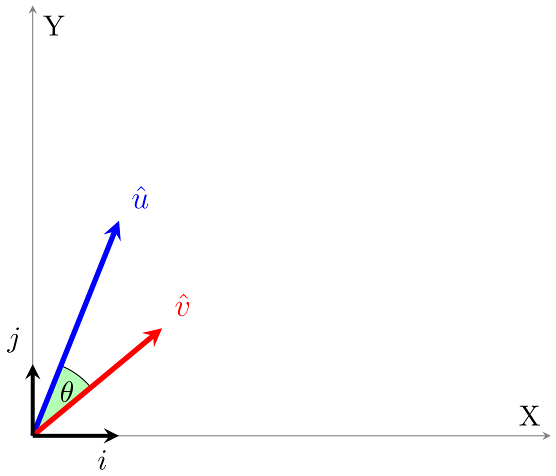


Figure 3.1 Two unit vectors

Thus, if

$$\hat{u} = u_1\hat{i} + u_2\hat{j}$$
$$\hat{v} = v_1\hat{i} + v_2\hat{j}$$

then, one should choose magnitudes, u_1, u_2, v_1, v_2 such that,

$$\|u\| = \sqrt{u_1^2 + u_2^2} = 1$$
$$\|v\| = \sqrt{v_1^2 + v_2^2} = 1$$

Using Pythagoras theorem, if we assume $\|u\| = K$, a constant, then as shown in figure 3.2,

$$u1 = K\text{cos}\beta$$
$$u2 = K\text{sin}\beta$$

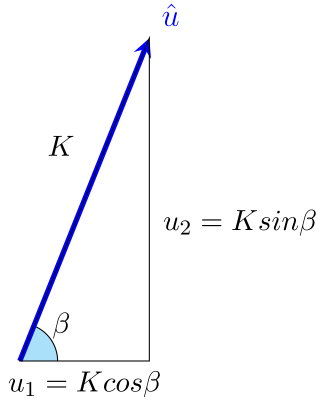


Figure 3.2 Magnitudes should add up to 1

$$\|u\| = \sqrt{K^2\text{cos}^2\beta + K^2\text{sin}^2\beta} = K = 1$$

Thus, we could conclude, for $\hat{u}$ to be unit vector,

$$u_1 = \text{cos}\beta$$
$$u_2 = \text{sin}\beta$$

We could similarly show that, if the angle spanned by $\hat{v}$ is α, then

$$v_1 = \text{cos}\alpha$$
$$v_2 = \text{sin}\alpha$$

Note that, $\theta = \beta - \alpha$ as shown in Figure 3.3.
According to Ptolemy's difference [1] from trignometry, one could write,

$$\text{cos}(\beta - \alpha) = \text{cos}\beta\text{cos}\alpha + \text{sin}\beta\text{sin}\alpha$$

Decomposing it as a *product matrix*,

$$\text{cos}(\beta - \alpha) = \begin{bmatrix}\text{cos}\beta & \text{sin}\beta\end{bmatrix}\begin{bmatrix}\text{cos}\alpha \\ \text{sin}\alpha\end{bmatrix}$$
$$\text{cos}\theta = \begin{bmatrix}u_1 & u_2\end{bmatrix}\begin{bmatrix}v_1 \\ v_2\end{bmatrix}$$

Whatever we are doing on the RHS above, we call *that* as **dot product** of vectors $\hat{u}, \hat{v}$. It is just that we *define* that quantity as a dot product, which is denoted by $\hat{u} \bullet \hat{v}$.

[1] https://www2.clarku.edu/faculty/djoyce/trig/ptolemy.html

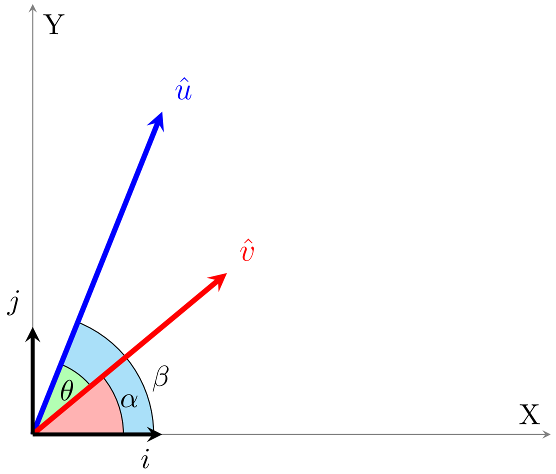


Figure 3.3: $\theta = \beta - \alpha$

Angle between two 2D unit vectors

$$\cos\theta = \hat{u} \bullet \hat{v} = \begin{bmatrix} u_1 & u_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = u_1v_1 + u_2v_2 \tag{3.1}$$

3.1.2 Angle between two 2D non unit vectors

Suppose we have *non unit* vectors, $\vec{a}, \vec{b}$ as shown in figure 3.4. We could derive their respective unit vectors easily by dividing with their magnitude.

Let

$$\vec{a} = a_1\hat{i} + a_2\hat{j}$$
$$\vec{b} = b_1\hat{i} + b_2\hat{j}$$

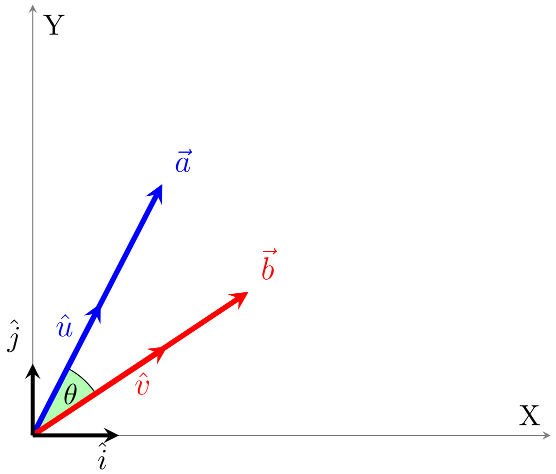


Figure 3.4: Dot product between non unit vectors $\vec{a}$ and $\vec{b}$

Then, their magnitudes will be,

$$\|a\| = \sqrt{a_1^2 + a_2^2}$$
$$\|b\| = \sqrt{b_1^2 + b_2^2}$$

The unit vectors could easily derived by scaling down to find unit x and y components

$$\hat{u} = \frac{a_1}{\|a\|}\hat{i} + \frac{a_2}{\|a\|}\hat{j}$$
$$\hat{v} = \frac{b_1}{\|b\|}\hat{i} + \frac{b_2}{\|b\|}\hat{j}$$

By using 3.1,

$$\begin{aligned}\cos\theta = \hat{u} \bullet \hat{v} &= \begin{bmatrix} \frac{a_1}{\|a\|} & \frac{a_2}{\|a\|} \end{bmatrix} \begin{bmatrix} \frac{b_1}{\|b\|} \\ \\ \frac{b_2}{\|b\|} \end{bmatrix} \\ &= \frac{a_1}{\|a\|}\frac{b_1}{\|b\|} + \frac{a_2}{\|a\|}\frac{b_2}{\|b\|} \\ &= \frac{a_1b_1 + a_2b_2}{\|a\|\|b\|}\end{aligned}$$

Taking $\|a\|\|b\|$ to the other side,

$$\begin{aligned}\|a\|\|b\|\cos\theta = a_1b_1 + a_2b_2 &= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= \vec{a} \bullet \vec{b}\end{aligned}$$

Thus,

$$\|a\|\|b\|\cos\theta = \vec{a} \bullet \vec{b}$$
$$\text{or} \quad \cos\theta = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|} \tag{3.2}$$

And we already have,

$$\vec{a} \bullet \vec{b} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = a_1b_1 + a_2b_2 \tag{3.3}$$

which is in *Matrix Multiplication* form. It is also conventional to write the same as in vector dot form as below.

$$\vec{a} \bullet \vec{b} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \bullet \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = a_1b_1 + a_2b_2 \tag{3.4}$$

Angle between two 2D non unit vectors

$$\cos\theta = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|}$$
$$\vec{a} \bullet \vec{b} = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \bullet \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = a_1b_1 + a_2b_2 \tag{3.5}$$

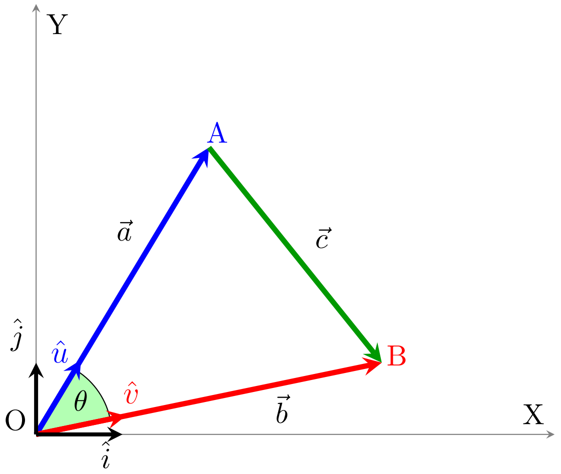


Figure 3.5: Dot Product
Setup for Alternate Proof

3.1.3 Law of Cosines

It is difficult to comprehend equation 3.1.1 in higher dimensions., so it could be helpful to try an alternate approach to derive the dot product. Suppose we have two vectors $\vec{a}, \vec{b}$. Then a 3rd vector $\vec{c}$ could be drawn making a triangle, such that, $\vec{a} + \vec{c} = \vec{b}$

Since $\vec{a} + \vec{c} = \vec{b}$, this implies, $\vec{c} = \vec{b} - \vec{a}$. Expanding,

$$\begin{aligned} c_1\hat{i} + c_2\hat{j} &= (b_1\hat{i} + b_2\hat{j}) - (a_1\hat{i} + a_2\hat{j}) \\ &= (b_1 - a_1)\hat{i} + (b_2 - a_2)\hat{j} \end{aligned} \tag{3.6}$$

Thus, their magnitudes also are equal.

$$\|\vec{c}\| = \sqrt{c_1^2 + c_2^2} = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2} = \|\vec{b} - \vec{a}\| \tag{3.7}$$

Let us draw perpendicular line from corners of the *triangle* and also have two more angles β, α defined as shown in figure 3.7.

Then,

$$\begin{aligned} \|\vec{a}\| &= \text{OA} \\ \|\vec{b}\| &= \text{OB} \\ \|\vec{c}\| &= \text{AB} = \text{AD} + \text{DB} \end{aligned}$$

By Pythagoras theorem, we could then say,

$$\begin{aligned} \text{AB} &= \text{AD} + \text{DB} = \text{OA}\text{cos}\beta + \text{OB}\text{cos}\alpha \\ \text{OA} &= \text{OE} + \text{EA} = \text{OB}\text{cos}\theta + \text{AB}\text{cos}\beta \\ \text{OB} &= \text{OF} + \text{FB} = \text{OA}\text{cos}\theta + \text{AB}\text{cos}\alpha \end{aligned}$$

Let $\text{AB} = c, \text{OA} = a, \text{OB} = b$, then

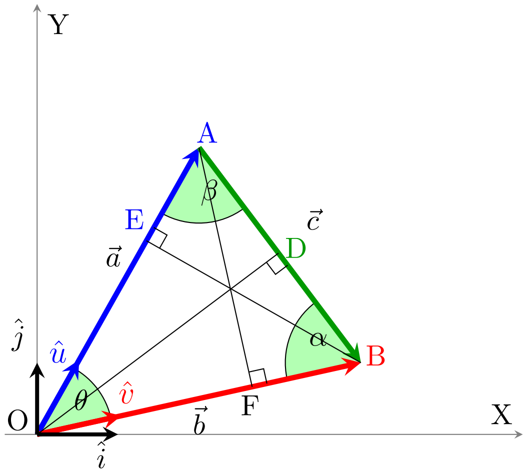


Figure 3.7: Introducing Perpendicular lines and two more angles α, β

$$\begin{aligned} c &= a\text{cos}\beta + b\text{cos}\alpha \\ a &= b\text{cos}\theta + c\text{cos}\beta \\ b &= a\text{cos}\theta + c\text{cos}\alpha \end{aligned}$$

Multiplying by the variable on LHS for all above three equations, we get,

$$\begin{aligned} c^2 &= ac\text{cos}\beta + cb\text{cos}\alpha \\ a^2 &= ab\text{cos}\theta + ac\text{cos}\beta \\ b^2 &= ab\text{cos}\theta + cb\text{cos}\alpha \end{aligned}$$

Combining as below,

$$\begin{aligned} (a^2 + b^2) - c^2 = ab\text{cos}\theta + \cancel{ac\text{cos}\beta} + ab\text{cos}\theta + \cancel{cb\text{cos}\alpha} \\ -\cancel{ac\text{cos}\beta} - \cancel{cb\text{cos}\alpha} = 2ab\text{cos}\theta \end{aligned}$$

Thus,

$$\begin{aligned} c^2 &= (a^2 + b^2) - 2ab\text{cos}\theta \\ \implies \text{AB}^2 &= (\text{OA}^2 + \text{OB}^2) - 2(OA)(OB)\text{cos}\theta \\ \implies \|c\|^2 &= \|a\|^2 + \|b\|^2 - 2\|a\|\|b\|\text{cos}\theta \end{aligned}$$

For simplicity, we shall use $c = \|c\|, a = \|a\|, b = \|b\|$ interchangeably.

Law of Cosines

The law of cosines states that, the lengths of the sides of a triangle could be related to cosine of one of its angles as below

$$\begin{aligned} c^2 &= (a^2 + b^2) - 2ab\text{cos}\theta \\ \|c\|^2 &= \|a\|^2 + \|b\|^2 - 2\|a\|\|b\|\text{cos}\theta \qquad (3.8) \end{aligned}$$

3.1.4 Angle between two 2D non unit vectors - Alternate Proof

Having established the law of cosines, we could then use that, to prove again the equation 3.5. Noting that,

$$\begin{aligned} a^2 &= a_1^2 + a_2^2 \\ b^2 &= b_1^2 + b_2^2 \end{aligned}$$

Also note, from equation 3.7,

$$c = b - a, \implies c^2 = (b-a)^2 = (b_1 - a_1)^2 + (b_2 - a_2)^2$$

Using both relations, we could thus re write equation 3.8 as,

$$\begin{aligned} 2ab\cos\theta &= (a^2 + b^2) - c^2 \\ &= (a_1^2 + a_2^2) + (b_1^2 + b_2^2) - (b_1 - a_1)^2 - (b_2 - a_2)^2 \\ &= (a_1^2 + a_2^2) + (b_1^2 + b_2^2) - (b_1^2 + a_1^2 - 2a_1b_1) - (b_2^2 + a_2^2 - 2a_2b_2) \\ &= \cancel{a_1^2} + \cancel{a_2^2} + \cancel{b_1^2} + \cancel{b_2^2} - \cancel{b_1^2} - \cancel{a_1^2} + 2a_1b_1 - \cancel{b_2^2} - \cancel{a_2^2} + 2a_2b_2 \\ &= 2(a_1b_1 + a_2b_2) \\ \therefore ab\cos\theta &= a_1b_1 + a_2b_2 \end{aligned}$$

Thus,

$$\cos\theta = \frac{a_1b_1 + a_2b_2}{ab} = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|}$$

which is same as equation 3.5.

3.1.5 Angle between two higher dimensional non unit vectors

As said earlier, law of cosines could be used similarly for higher dimensions. This is because, at any higher dimension, the angle between two vectors is still on a plane that contains the two vectors, thus on that plane, the relation we just saw, apply. Figure 3.7 illustrates the case for 3 dimensions.

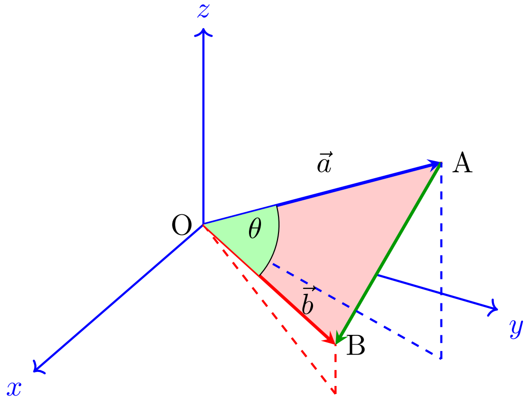


Figure 3.7: Law of Cosine applicable in any $n > 1$ dimensions

Thus, if we define 3 dimensional vectors as below

$$\begin{aligned}
\vec{a} &= a_1\hat{i} + a_2\hat{j} + a_3\hat{k} = \langle a_1, a_2, a_3 \rangle \\
\vec{b} &= b_1\hat{i} + b_2\hat{j} + b_3\hat{k} = \langle b_1, b_2, b_3 \rangle \\
\vec{c} = \vec{b} - \vec{a} &= (b_1 - a_1)\hat{i} + (b_2 - a_2)\hat{j} + (b_3 - a3)\hat{k}
\end{aligned}$$

then using 3.8,

$$\begin{aligned}
2ab\cos\theta &= (a^2 + b^2) - c^2 \\
&= (a_1^2 + a_2^2 + a_3^2) + (b_1^2 + b_2^2 + b_3^2) - [(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2] \\
&= (a_1^2 + a_2^2 + a_3^2) + (b_1^2 + b_2^2 + b_3^2) - [b_1^2 + a_1^2 - 2a_1b_1 + b_2^2 + a_2^2 - 2a_2b_2 + b_3^2 + a_3^2 - 2a_3b_3 \\
&= \cancel{(a_1^2 + a_2^2 + a_3^2)} + \cancel{(b_1^2 + b_2^2 + b_3^2)} - [\cancel{(b_1^2 + b_2^2 + b_3^2)} + \cancel{(a_1^2 + a_2^2 + a_3^2)} - 2a_1b_1 - 2a_2b_2 - 2a_3b_3] \\
&= 2(a_1b_1 + a_2b_2 + a_3b_3) \\
\therefore ab\cos\theta &= a_1b_1 + a_2b_2 + a_3b_3
\end{aligned}$$

Thus,

$$\cos\theta = \frac{a_1b_1 + a_2b_2 + a_3b_3}{ab} = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|}$$

which is same as equation 3.5.

In fact, we could also prove for any n dimensional vector as below even if we are unable to visualize beyond 3D. Note the short form used to denote the vector. If we define n dimensional vectors as below

$$\begin{aligned}
\vec{a} &= \langle a_1, a_2, a_3, \cdots, a_n \rangle \\
\vec{b} &= \langle b_1, b_2, b_3, \cdots, b_n \rangle \\
\vec{c} = \vec{b} - \vec{a} &= \langle (b_1 - a_1), (b_2 - a_2), \cdots, (b_n - a_n) \rangle
\end{aligned}$$

then using 3.8,

$$\begin{aligned}
2ab\cos\theta &= (a^2 + b^2) - c^2 \\
&= (a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2) + (b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2) - \\
&[(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2 + \cdots + (b_n - a_n)^2] \\
&= (a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2) + (b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2) - \\
&[b_1^2 + a_1^2 - 2a_1b_1 + b_2^2 + a_2^2 - 2a_2b_2 + \cdots + b_n^2 + a_n^2 - 2a_nb_n] \\
&= \cancel{(a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)} + \cancel{(b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2)} - \\
&[\cancel{(b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2)} + \cancel{(a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2)} - \\
&(2a_1b_1 + 2a_2b_2 + \cdots + 2a_nb_n)] \\
&= 2(a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n) \\
\therefore ab\cos\theta &= a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n
\end{aligned}$$

Thus,

$$\cos\theta = \frac{a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n}{ab} = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|}$$

Angle between two any n dimensional non unit vectors

For any two $n > 1$ dimensional vectors, the angle between them could always be calculated as,

$$\cos\theta = \frac{a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n}{ab} = \frac{\vec{a} \bullet \vec{b}}{\|a\|\|b\|} \tag{3.9}$$

This is true, even when we are unable to comprehend visually beyond 3D vectors, because the angle between two vectors is always on a plane (2D) no matter the dimensionality of the vectors is.

Bibliography

[1] J. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole CENGAGE Learning, 8th edition, 2011. URL https://fac.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf.

[2] J. Frost. Heteroscedasticity in regression analysis. 2017. URL http://statisticsbyjim.com/regression/heteroscedasticity-regression/.

[3] Robert, Elliot, and Dale. *Probability and Statistical Inference*. Pearson, 9th edition, 2015. URL http://www.nylxs.com/docs/thesis/sources/Probability%20and%20Statistical%20Inference%209ed%20%5B2015%5D.pdf.

[4] C. Zaiontz. Basic concepts of correlation. 2013. URL http://www.real-statistics.com/correlation/basic-concepts-correlation/.

[5] Y. Zhang, L. Cheng, and H. Wu. Some new deformation formulas about variance and covariance. 2012. URL https://www.researchgate.net/profile/Yuli_Zhang/publication/261496020_Some_new_deformation_formulas_about_variance_and_covariance/links/54eda4c80cf25da9f7f1274e/Some-new-deformation-formulas-about-variance-and-covariance.pdf.