3.5 Measures of Association Between 2 variables

1 Covariance and 2 Correlation are descriptive measures for the relationship between two variables.

---

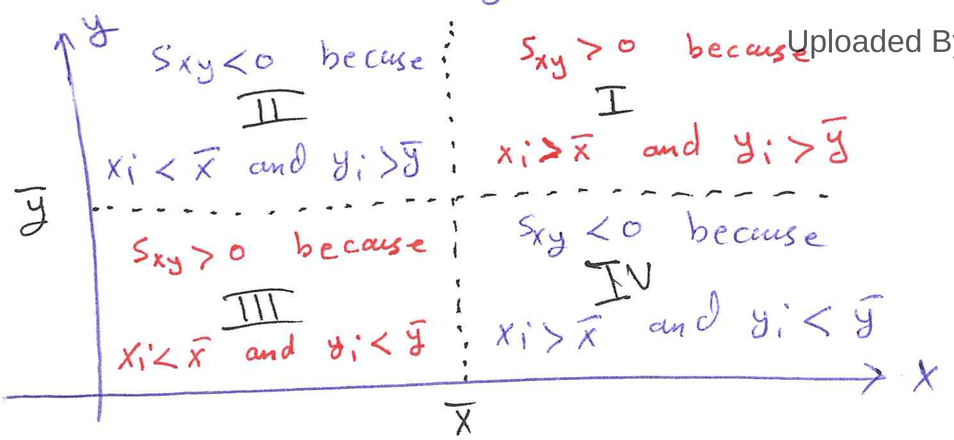
1 **Covariance:** • For a sample of size n with observations (x_i, y_i), $i = 1, 2, \dots, n$

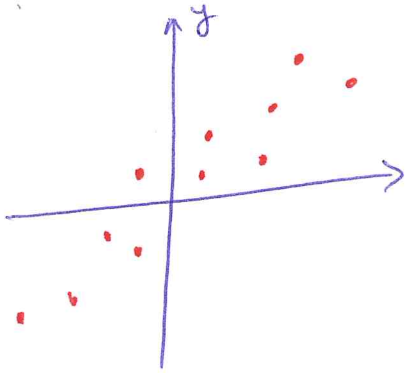
the sample covariance is $s_{xy} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$

• Population covariance $\sigma_{xy} = \dfrac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$

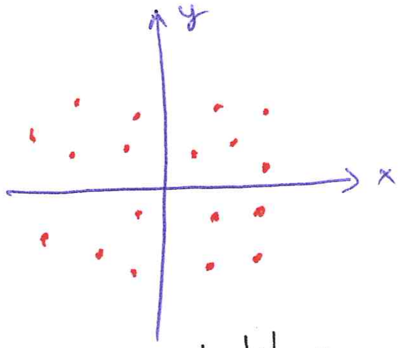
• Interpretation of the covariance:- The covariance is a measure of the linear association between two variables.

- If $s_{xy} > 0$, then there is a positive linear association between x and y. (If $x\uparrow$, then $y\uparrow$)
- If $s_{xy} < 0$, then there is a negative linear association between x and y. (If $x\uparrow$, then $y\downarrow$)
- If $s_{xy} = 0$ "or close to zero", then there is no linear association between x and y.

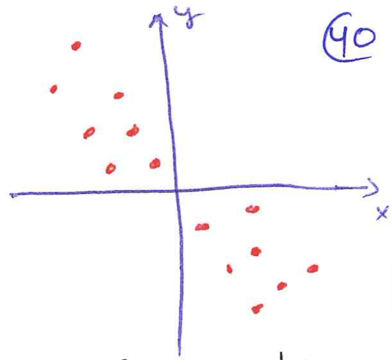




S_{xy} Positive
(x and y are positively linearly related)



S_{xy} approximately 0
(x and y are not linearly related)

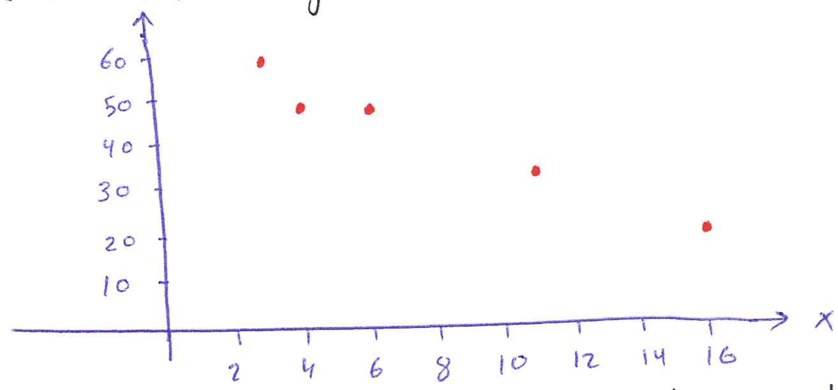


S_{xy} Negative
(x and y are negatively linearly related)

Example: (Q45) Five observations taken for two variables:

x_i	4	6	11	3	16
y_i	50	50	40	60	30

[a] Develop a scatter diagram with x on the horizontal axis.



[b] What does the scatter diagram indicate about the relationship between the two variables?

There is a negative linear relationship between x and y

[c] Compute and interpret the sample covariance.

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
4	50	-4	4	-16
6	50	-2	4	-8
11	40	3	-6	-18
3	60	-5	14	-70
16	30	8	-16	-128
40	230			-240

$\bar{x} = \frac{\sum x_i}{n} = \frac{40}{5} = 8$

$\bar{y} = \frac{\sum y_i}{n} = \frac{230}{5} = 46$

$S_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n-1}$

$= \frac{-240}{4} = -60$

The sample covariance indicate a negative linear association between x and y.

[2] Correlation Coefficient

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$, where

r_{xy} = sample correlation coefficient "pearson product moment correlation coefficient"

S_{xy} = sample covariance

S_x = sample standard deviation of x $\qquad s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$

S_y = Sample standard deviation of y. $\qquad s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$

---

* Pearson product moment correlation coefficient "Population data"

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$, where

ρ_{xy} = population correlation coefficient

σ_{xy} = population covariance

σ_x = population standard deviation for x

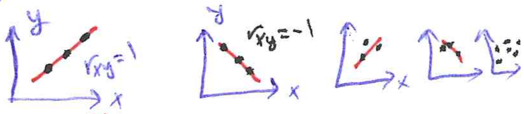
σ_y = population standard deviation for y

---

* The sample correlation coefficient r_{xy} is the estimator of the population correlation coefficient ρ_{xy}.

Interpretation of the correlation coefficient:

- The correlation coefficient ranges from -1 to 1 $\qquad r_{xy} \in [-1, 1]$
- When the " " values close to -1, it indicate a strong linear negative relationship between x and y.
- " " " " " " " " 1, " " " " " positive " " " " " " "
- " " " " " " " " 0 " " " " a weak relaship between x and y.



scatters

⇒ Example (Q45)

d) Compute and interpret the sample correlation coefficient.

x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$
4	50	-4	4	16	16
6	50	-2	4	4	16
11	40	3	-6	9	36
3	60	-5	14	25	196
16	30	8	-16	64	256
				118	520

$$S_x = \sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{118}{5-1}} = \sqrt{29.5} = 5.43$$

$$S_y = \sqrt{\frac{\Sigma(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{520}{5-1}} = 11.4$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-60}{(5.43)(11.4)} = -0.969$$

The sample correlation coefficient (-0.969) indicates of a stronge negative linear relationship.

