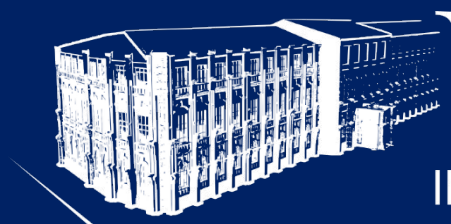


WESTMINSTER

INTERNATIONAL UNIVERSITY IN TASHKENT

An Accredited Institution of the University of Westminster (UK)

LECTURE 8
CORRELATION AND REGRESSION

Saidg'ozi Saydumarov
Sherzodbek Safarov
QM Module Leaders
ssaydumarov@wiut.uz
s.safarov@wiut.uz

Office hours: by appointment
Room ATB308
EXT: 660

Lecture outline

- Quick review
- Covariance
- Correlation
- Regression

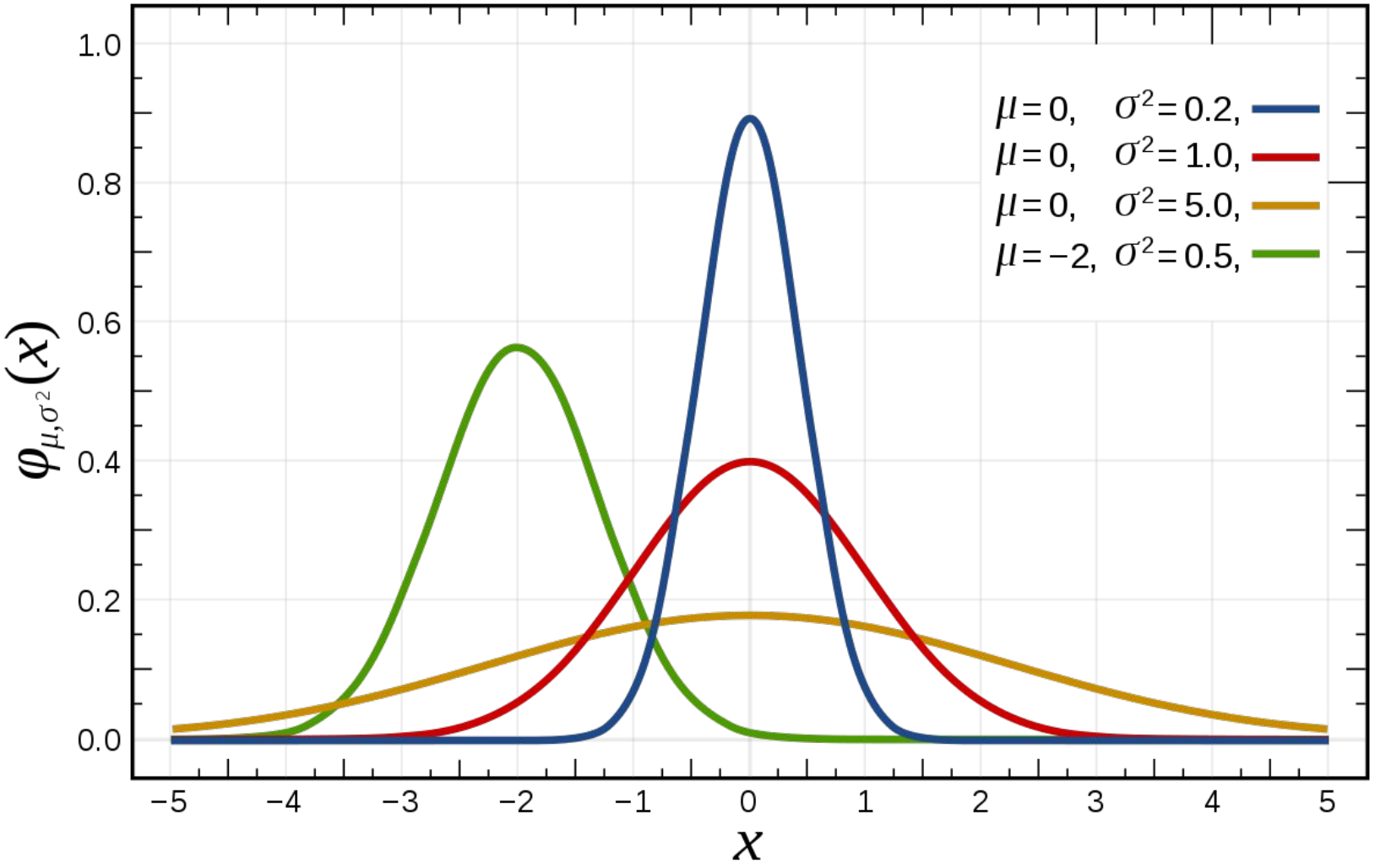
Quick review

- ***Population*** ***Sample***

$$\sigma^2 = \sum \frac{(x_i - \bar{x})^2}{n} \qquad \sigma^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

$$\sigma = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n}} \qquad \sigma = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}$$

Quick review



Covariance

- ***X set (Humidity)***

$$\sigma_x = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}} = 12.59$$

Time of Year	Humidity
January	73%
February	68%
March	62%
April	60%
May	53%
June	40%
July	39%
August	42%
September	45%
October	57%
November	66%
December	73%

- ***Y set (Sales of umbrella)***

$$\sigma_y = \sqrt{\sum \frac{(y_i - \bar{y})^2}{n-1}} = 0.96$$

Time of Year	Sales of umbrella
January	2.50
February	2.20
March	2.10
April	2.00
May	1.50
June	0.50
July	0.01
August	0.05
September	0.30
October	1.20
November	1.80
December	2.60

Covariance

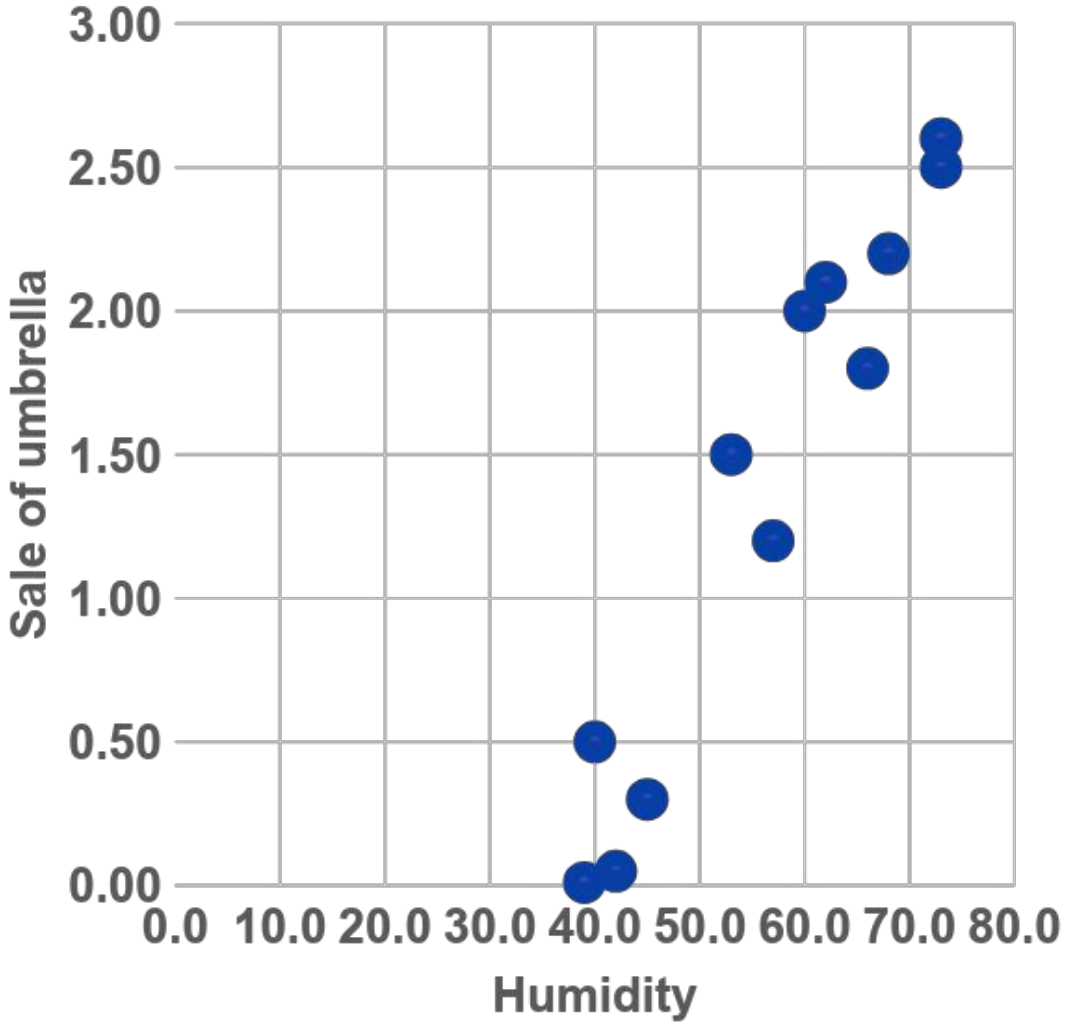
Population covariance

$$Cov_{x,y} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{n}$$

Sample covariance

$$Cov_{x,y} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{n - 1}$$

Covariance *measures how two variables move with respect to each other and is an extension of the concept of variance (which tells about how a single variable varies)*



Covariance

Population covariance

$$Cov_{x,y} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{n}$$

Sample covariance

$$Cov_{x,y} = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{n - 1}$$

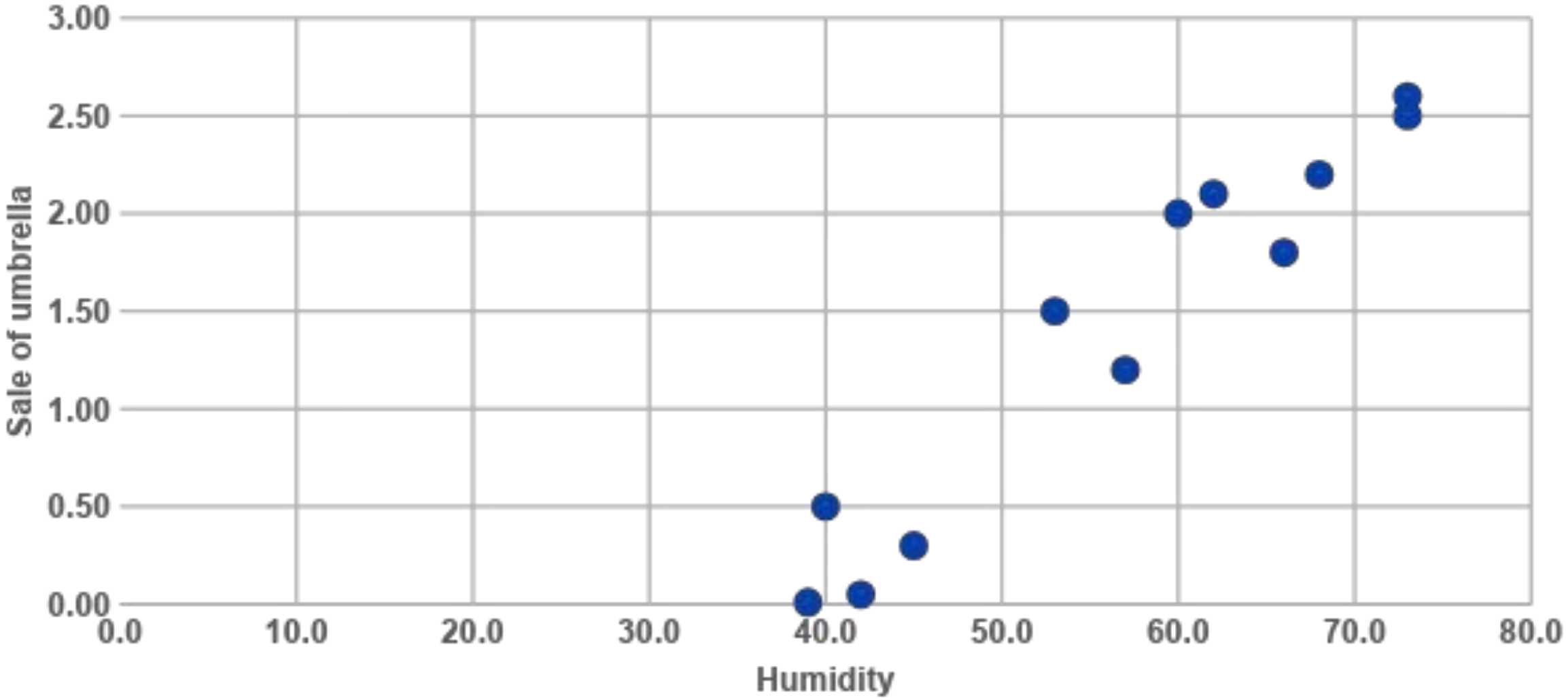
Covariance *measures how two variables move with respect to each other and is an extension of the concept of variance (which tells about how a single variable varies)*

	X (humidity)	Y (sales of umbrella)			
	73	2.50	16.5	1.1	18.2
	68	2.20	11.5	0.8	9.2
	62	2.10	5.5	0.7	3.9
	60	2.00	3.5	0.6	2.1
	53	1.50	-3.5	0.1	-0.4
	40	0.50	-16.5	-0.9	14.8
	39	0.01	-17.5	-1.4	24.3
	42	0.05	-14.5	-1.3	19.5
	45	0.30	-11.5	-1.1	12.6
	57	1.20	0.5	-0.2	-0.1
	66	1.80	9.5	0.4	3.8
	73	2.60	16.5	1.2	19.9
Sum	**678**	**16.76**			**127.9**
Mean	**56.5**	**1.40**			

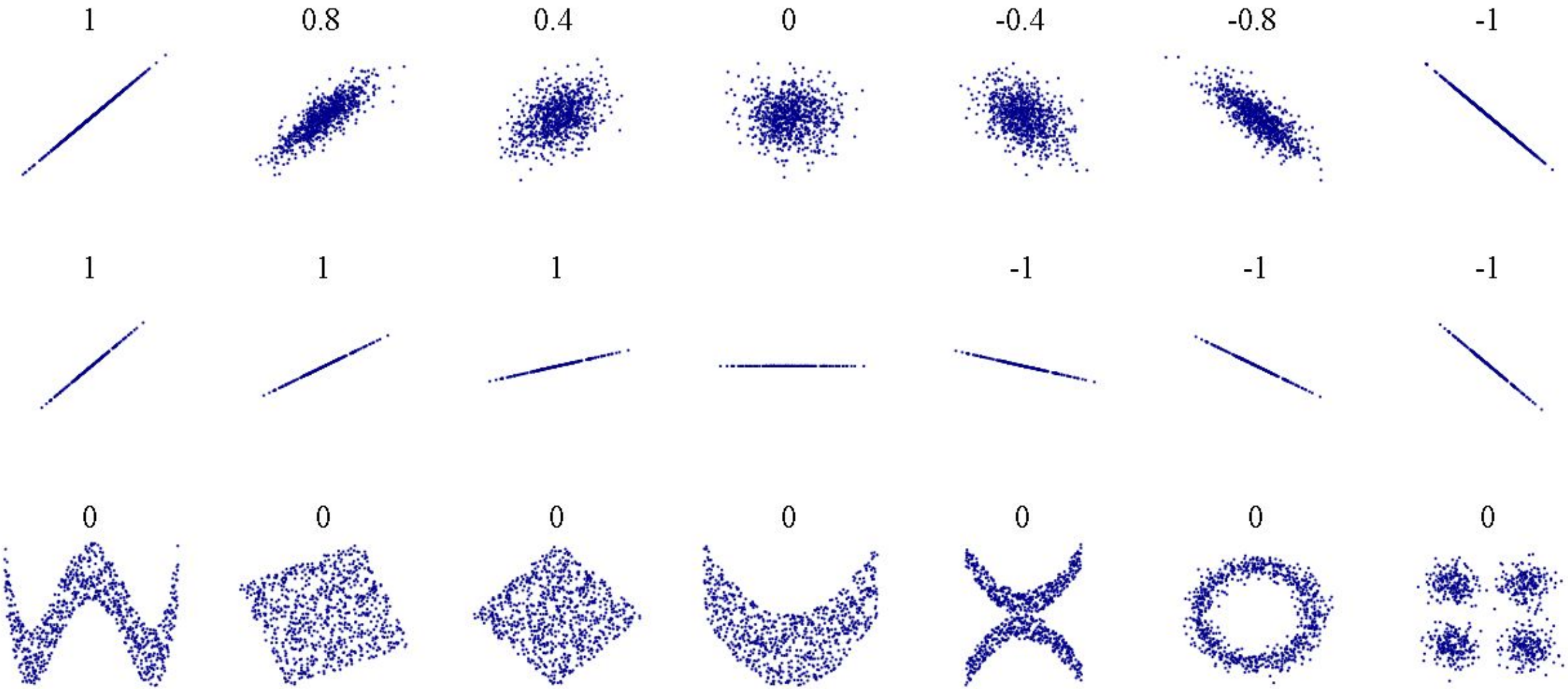
$$Cov(x,y)_{population} = \frac{127.9}{12} = 10.7$$

$$Cov(x,y)_{sample} = \frac{127.9}{11} = 11.6$$

Correlation



Correlation



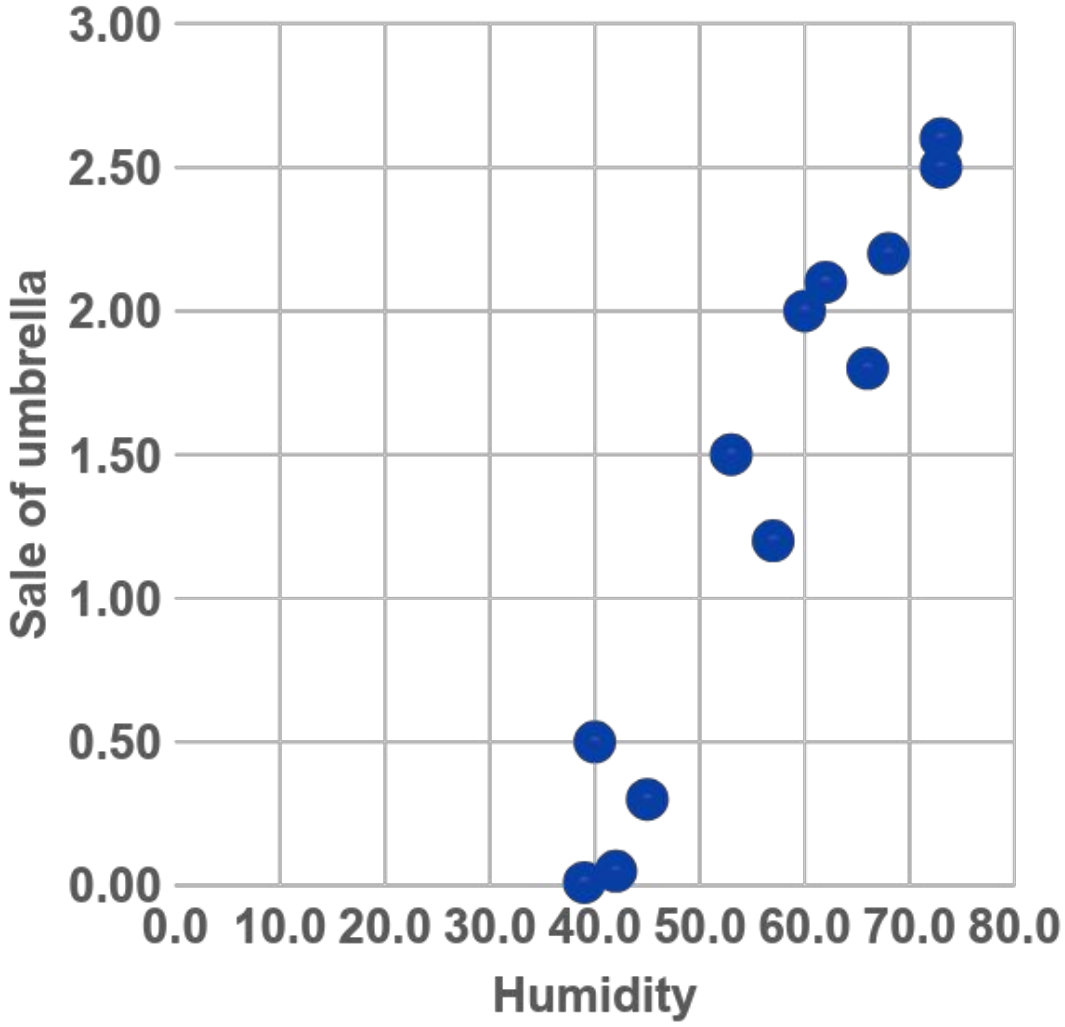
Correlation

- $$\rho_{x,y} = \frac{cov_{x,y}}{\sigma_x \sigma_y}$$

Correlation *quantifies the relationship between two random variables. In simple terms, it is a unit measure of how these variables change with respect to each other (normalized covariance value).*

$$\rho_{x,y} = \frac{cov_{x,y}}{\sigma_x \sigma_y} = \frac{11.6}{12.59 \times 0.96} = 0.9624 = 96.24\%$$

It can only take values between +1 (100%) and -1 (-100%). A correlation of +1 (100%) indicates that random variables have a direct and strong relationship.



Covariance vs Correlation

- ***Covariance***

- Shows how strong two variables are related to each other given constant condition
- Unlimited, hence lies in the range of $+\infty$ and $-\infty$
- Rescaling and other adjustments affects

Correlation

- Shows the extent to which two variables are dependent on each other.
- Limited to values between +1 and -1
- Not effected if rescaled or multiplied by a factor change

Regression

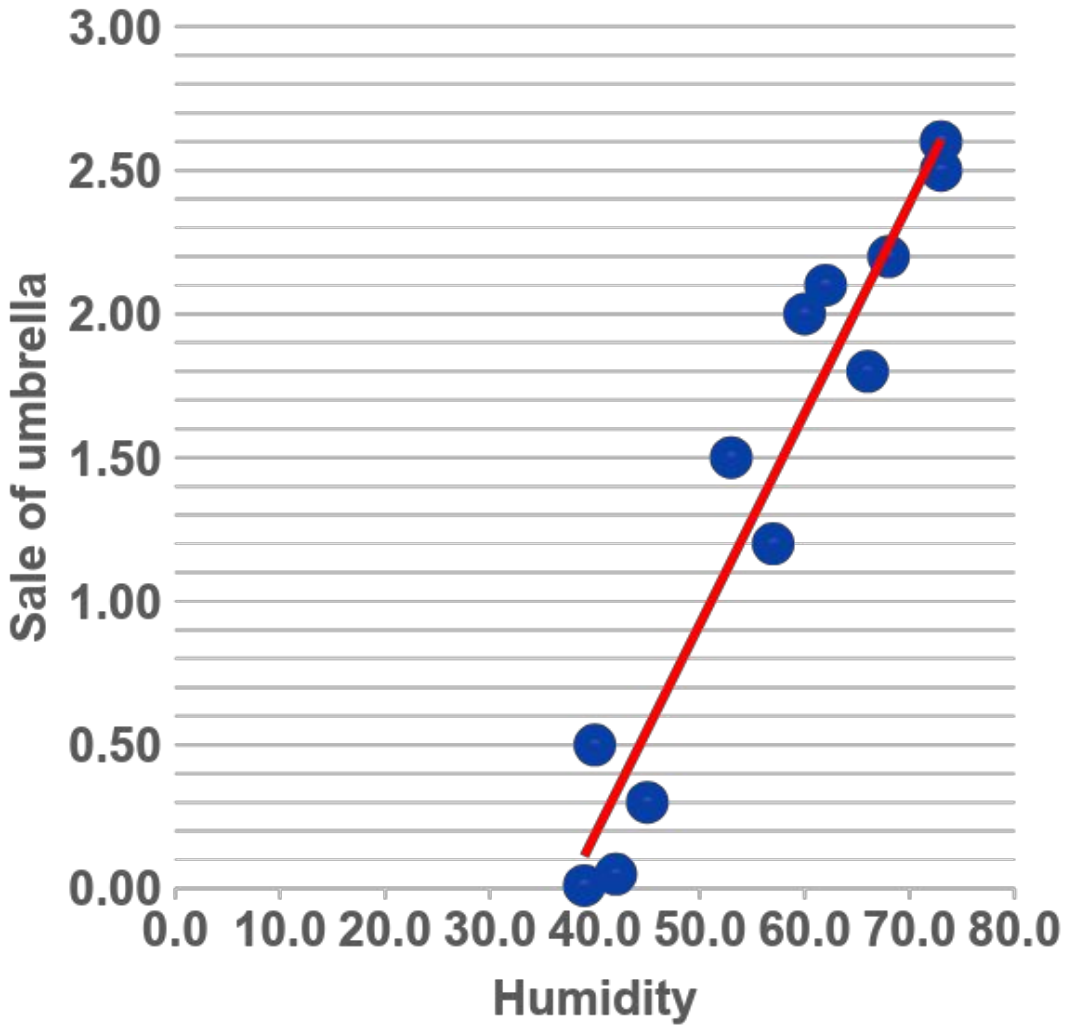
Regression *is a technique for determining the statistical relationship between two or more variables where a change in a dependent variable is associated with, and depends on, a change in one or more independent variables.*

$$y = a + bx$$

$$y_i = \alpha + \beta x_i + \varepsilon$$

$$\boldsymbol{\beta} = \frac{\boldsymbol{cov}_{x,y}}{\boldsymbol{\sigma}_x^2} = \boldsymbol{\rho}_{x,y} \times \frac{\boldsymbol{\sigma}_y}{\boldsymbol{\sigma}_x} = \frac{11.62}{12.59^2}$$
$$= 0.9624 \times \frac{0.96}{12.59} = 0.073$$

$$\boldsymbol{\alpha} = \bar{\boldsymbol{y}} - \boldsymbol{\beta}\bar{\boldsymbol{x}} = 1.40 - 0.073 \times 56.50 = -2.725$$

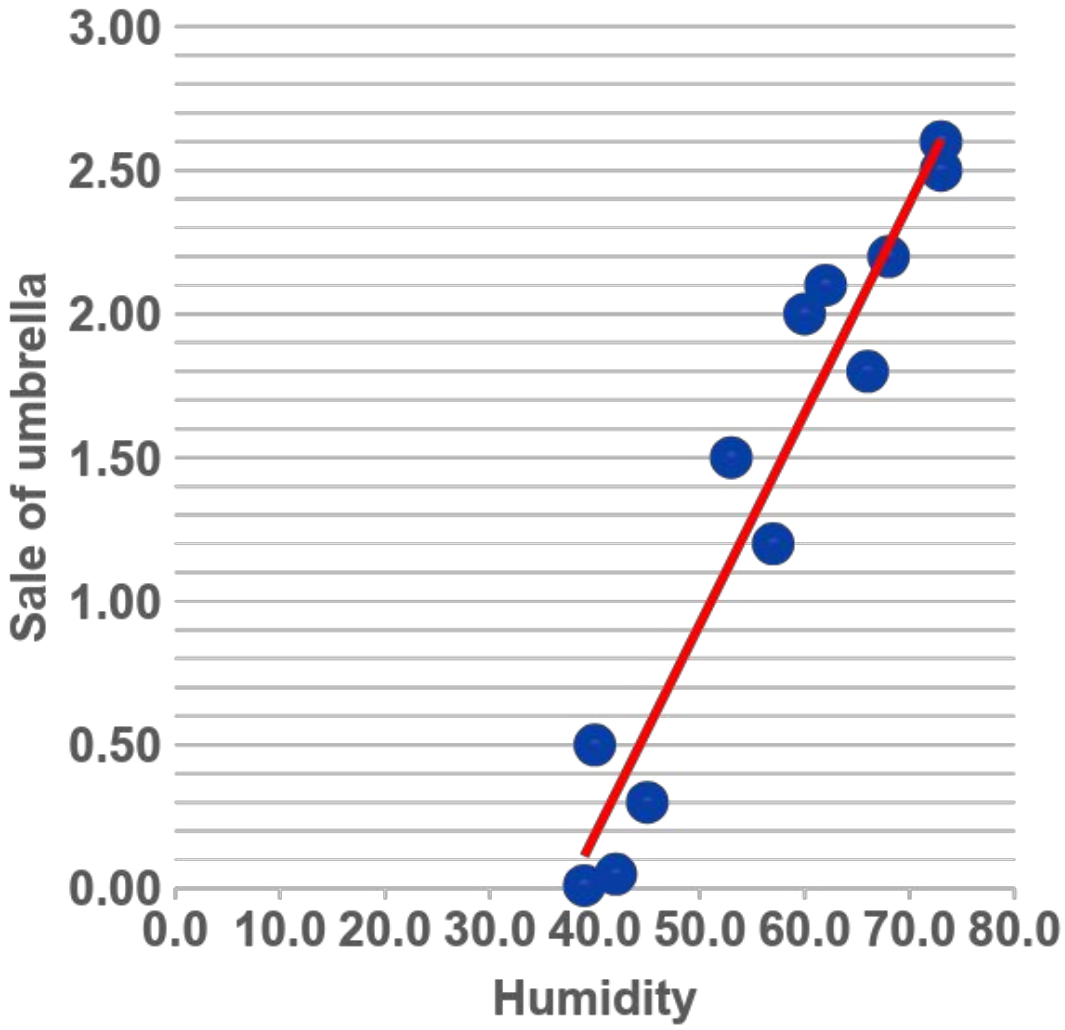


Regression

R-squared *is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. In other words, R-squared simply measures to what extend independent variable, in terms of variances, explains the dependent variable.*

It takes values only between 0 and 1, which is the same as 0% and 100%, respectively.

$$R^2 = \rho_{x,y}^2 = 0.9624^2 = 0.9262$$




Thank
you