Covariance and Correlation

Dr. Linta Rose

rose.l@incois.gov.in

Covariance

- A systematic relationship between a pair of random variables, wherein a change in one variable is reciprocated by an equivalent change in the other
- Covariance can take any value between $-\infty$ and $+\infty$, where the negative value is an indicator of negative relationship, wheras a positive value represents positive relationship and when the value is zero, it indicates no relationship.
- In addition to this, when all the observations of either variable are same, the covariance is zero.
- When we change the unit or scale of any or both the variables, then there is no change in the strength of the relationship between the variable, but the value of covariance changes.

Covariance

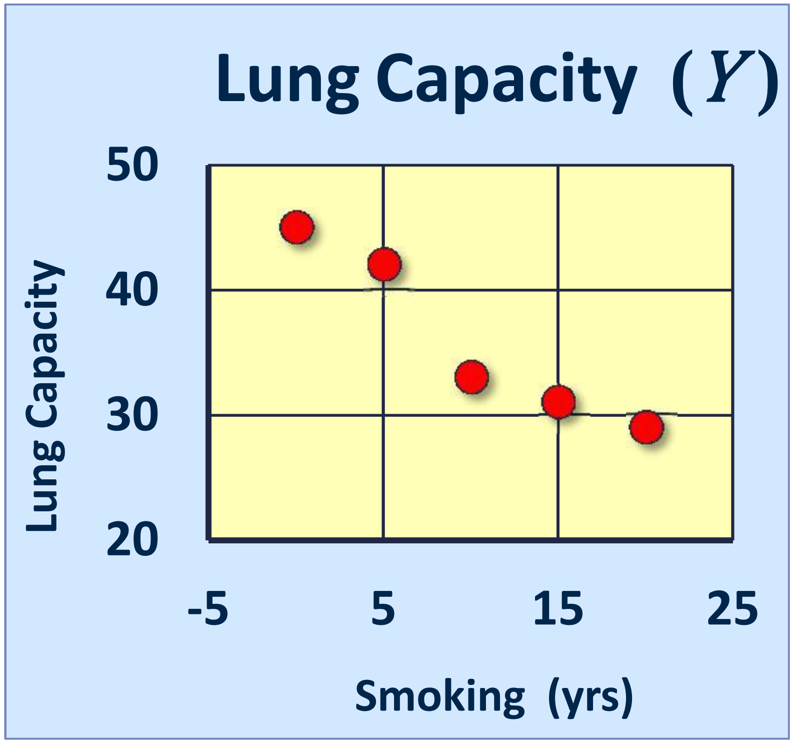
- Variables may change in relation to each other. Covariance is a measure of association of two variables.
- If positive, then both variables increase or decrease together. If negative, then they vary in opposite manner.

- *Covariance* measures how much the movement in one variable predicts the movement in a corresponding variable
- Example: investigate relationship between *cigarette smoking* and *lung capacity*

Smoking vs Lung Capacity

N	Cigarettes (X)	Lung Capacity (Y)
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

- Variables smoking and lung capacity *covary* inversely, like



Covariance

- Average *product of deviation* measures extent to which variables co-vary, the degree of linkage between them
- Similar to variance, for theoretical reasons, average is typically computed using (N-1), not N. Thus, covariance

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\underbrace{\left(X_i - \bar{X}\right)}_{\text{Deviation of data 1 from mean}}\underbrace{\left(Y_i - \bar{Y}\right)}_{\text{Deviation from mean of data 2}}$$

Calculating Covariance

Cigs (X)				Cap (Y)
0	-10	-90	9	45
5	-5	-30	6	42
10	0	0	-3	33
15	5	-25	-5	31
20	10	-70	-7	29
		$\sum$= -215		

Evaluation yields,

$$S_{xy} = \frac{1}{4}(-215) = -53.75$$

Correlation

- A measure which determines the change in one variable due to change in other variable.

- Correlation is of two types, i.e. positive correlation or negative correlation.

- Correlation can take any value between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 is an indicator of strong negative correlation.

- There are four measures of correlation:
  - Scatter diagram
  - Rank correlation coefficient

Correlation coefficient

- Covariance, Cov(X,Y) is dependent upon the units of X & Y.
- Correlation, Corr(X,Y), scales covariance by the standard deviations of X & Y so that it lies between 1 & –1

$$Corr(x, y) = \frac{Cov(x,y)}{\pm_x \pm_y}$$

Where ± is the Standard deviation

Correlation Analysis

A statistical analysis used to obtain a quantitative measure of the strength of the linear relationship between a dependent variable and one or more independent variables

Correlation coefficient formula

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

This can be rearranged as

$$r_{xy} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left(N\sum X^2 - \left(\sum X\right)^2\right)\left(N\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

Table for Calculating *corr*

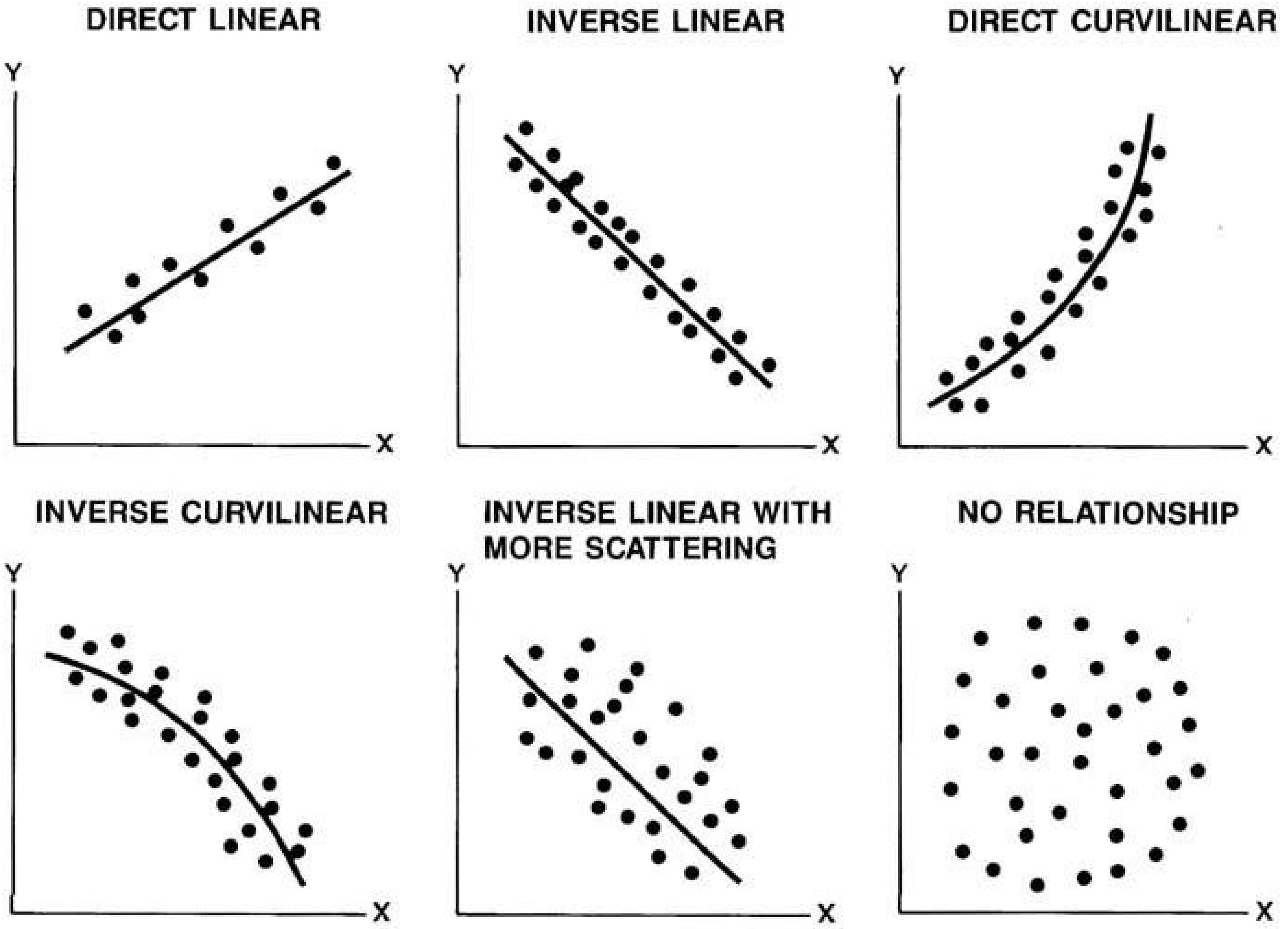
$$\boxed{r_{xy} = -0.9615}$$

	Cigs (X)	X^2	XY	Y^2	Cap (Y)
	0	0	0	2025	45
	5	25	210	1764	42
	10	100	330	1089	33
	15	225	465	961	31
	20	400	580	841	29
$\sum=$	50	750	1585	6680	180

- r_{xy} = -0.96 implies almost certainty smoker will have diminish lung capacity
- Greater smoking exposure implies greater likelihood of lung damage

Correlation – Scatter Diagram

Visual Relationship Between X and Y



Key Differences

Key	Covariance	Correlation
Meaning	Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.
What is it?	Measure of correlation	Scaled version of covariance
Values	Lie between -∞ and +∞	Lie between -1 and +1
Change in scale	Affects covariance	Does not affects correlation
Unit free measure	No	Yes
Significance	provides direction of relationship	provides direction and strength of relationship

Summary

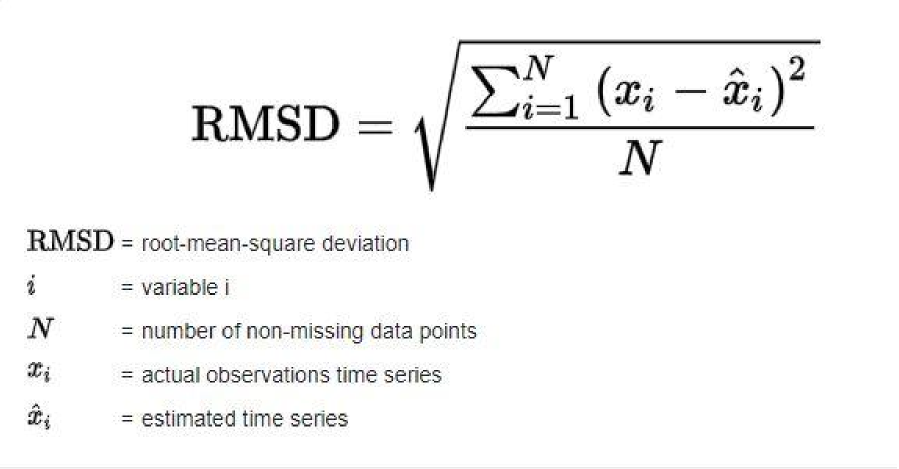
- Correlation is a special case of covariance which can be obtained when the data is standardized.

- Now, when it comes to making a choice, which is a better measure of the relationship between two variables, correlation is preferred over covariance, because it remains unaffected by the change in location and scale, and can also be used to make a comparison between two pairs of variables.

Bias and RMSE

- Measures of goodness for comparison of two variables – eg. model vs. observations
- Bias – how close is the estimated value to the true value

- RMSE – Root Mean Square of Error – used mostly in regression
  - To check the quality of prediction errors



$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

RMSD	= root-mean-square deviation
i	= variable i
N	= number of non-missing data points
x_i	= actual observations time series
$\hat{x}_i$	= estimated time series