5. Bivariate Analysis – Correlation and Regression

Dr. Prasad Modak

Environmental Management Centre, Mumbai

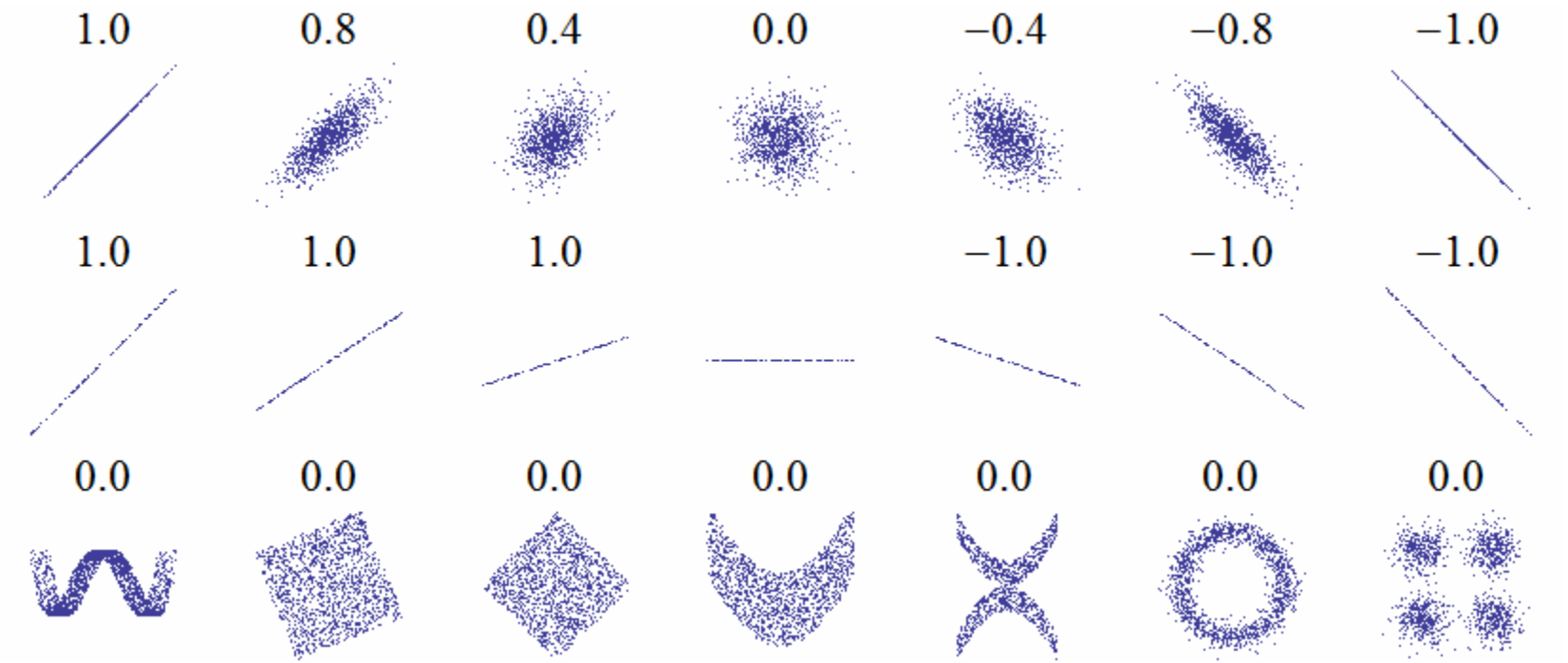
Correlation

- Indicates the strength and direction of a linear relationship between two independent variables
-
- Pearson's product-moment correlation

$$r_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x \sigma_y}$$

- Correlation coefficient's (r) values always range from -1 to +1
- - indicates negative, and + indicates positive correlation
- The closed the r value to r the higher the correlation

Example of correlation between any X & Y



Correlation	Negative	Positive
None	−0.09 to 0.0	0.0 to 0.09
Small	−0.3 to −0.1	0.1 to 0.3
Medium	−0.5 to −0.3	0.3 to 0.5
Large	−1.0 to −0.5	0.5 to 1.0

http://upload.wikimedia.org/wikipedia/commons/0/02/Correlation_examples.png

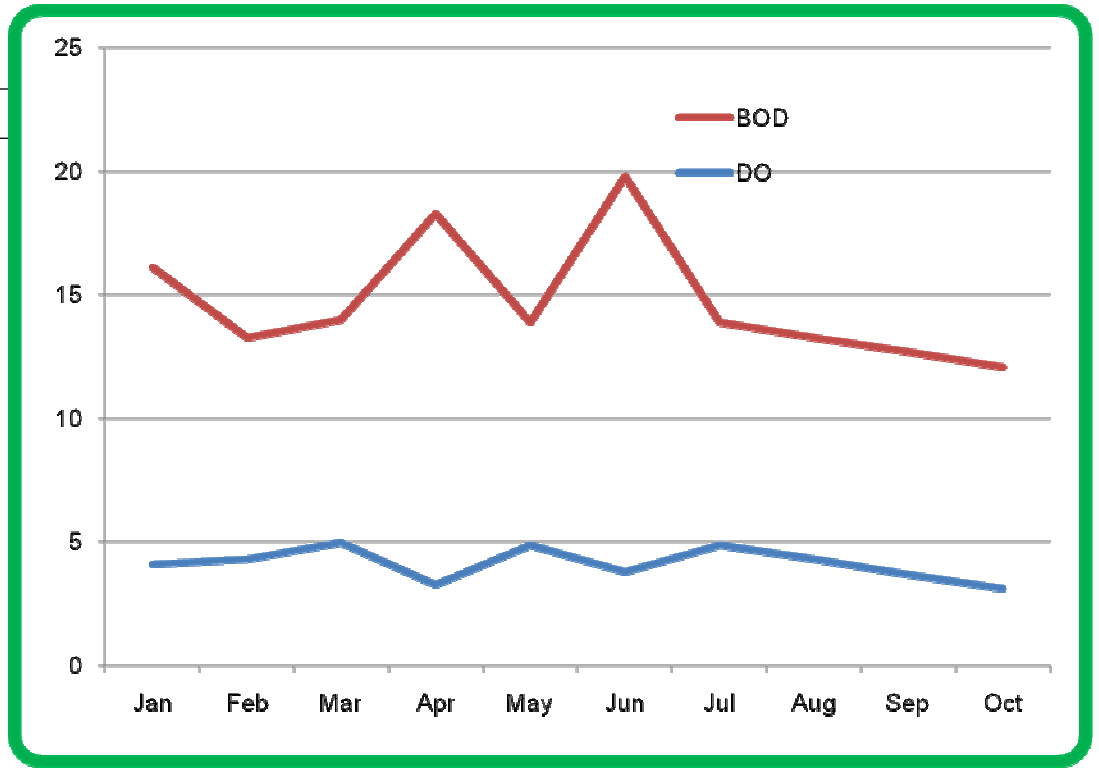
Example

Sea water at Nariman Point, 2008

Month	DO	BOD
Jan	4.1	12
Feb	4.3	9
Mar	5	9
Apr	3.3	15
May	4.9	9
Jun	3.8	16
Jul	4.9	9
Aug	4.3	9
Sep	3.7	9
Oct	3.1	9

http://mpcb.gov.in/envtdata/wqwebpg.php?rgnId=12

r ***-0.45335***



Regression

- Regression provides an estimate of one variable from another using a mathematical expression
- Shows relationship between a dependent variable and one or more independent variables
- it is a simple tool to
  - Find out missing values
  - Perform basic environmental modeling
- A best fit line is drawn with the eqn. $y = \alpha x + \beta$
- Where :

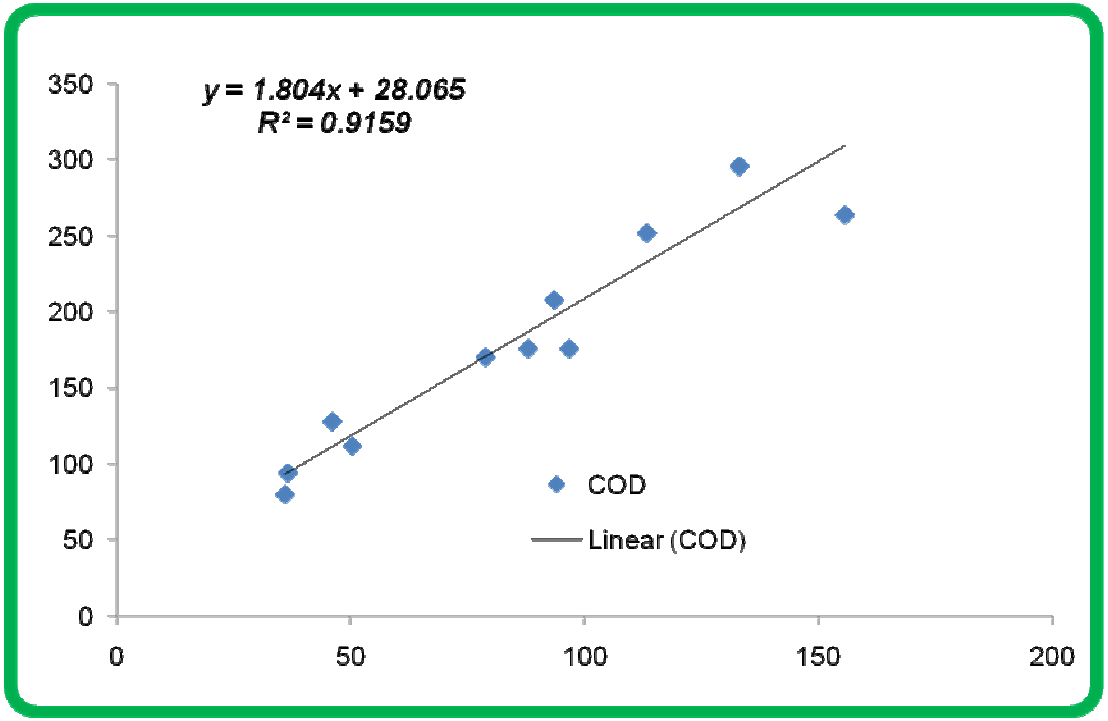
$$\beta = \frac{n\sum xy - (\sum x.\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\alpha = \frac{\sum y - \beta \sum x}{n}$$

Example

Month	BOD	COD
Jan	113.4	252
Feb	155.76	264
Mar	133.2	296
Apr	93.6	208
May	88	176
Jun	133.2	296
Jul	96.8	176
Aug	36	80
Sep	46.08	128
Oct	50.4	112
Nov	78.9	170.4
Dec	36.6	94.09

COD (dependent var.) is 'described' with the help of BOD (independent var.)



- So for a missing value of dependent variable (COD) we could use this regression equation & corresponding BOD values
- The regression equation is essentially a 'model'
- R^2 is coefficient of determination

Coefficient of Determination

- Coefficient of determination, R^2 is the proportion of variability in a data set that is accounted for by the statistical model
- R^2 values vary from 0 to 1
- R^2 gives extent of variance explained by the model.
- R^2 of 0.6 implies that 60% of the variance of the dependent variable is explained by the regression model.
- Higher is R^2 better is the regression model