Correlation

An **association** exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.

A **scatter plot** is a graphical display for two quantitative variables. One variable is denoted by '*x*' and represented along the horizontal axis. The other variable is denoted by '*y*' and represented along the vertical axis. The values for *x* and *y* for an individual or object is represented by a point relative to the two axis.

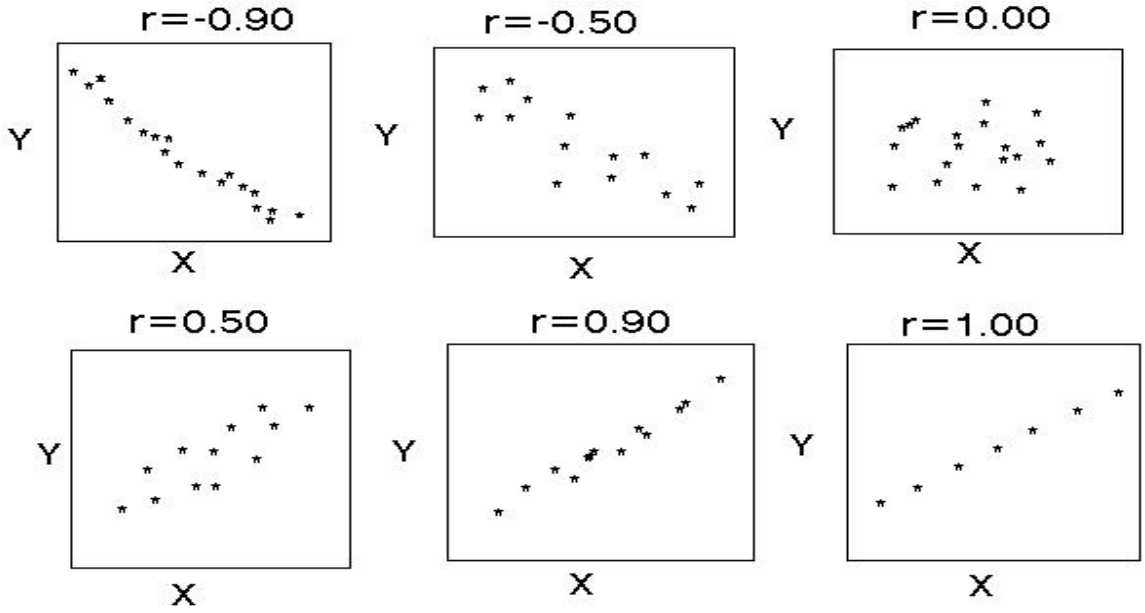
Positive association: As *x* increases *y* also tend to increase. Both variables behave in the same direction.

Negative association: As *x* increases *y* tend to decrease. Variables behave in the opposite direction.

The measure **correlation** summarizes the direction of the association between two quantitative variables as well as the strength of their straight-line trend. The symbol for sample correlation is r and the population symbol for correlation is **ρ (rho)**.

Properties of correlation

1) It takes values between **-1** and **+1**.
2) The closer r is to ± 1, the closer the data points fall to a straight line and the **stronger** is the linear association.
3) The closer r is to 0, the **weaker** is the linear association.



Regression Analysis:

Regression analysis describes the relationship between two numerical variables in a specific setting. Of the two variables, the 'variable of interest' in a study is known as the "**dependent**" variable (y) and the other variable is called the "**independent**" variable (x) that explains changes in the dependent variable.

The **regression line** predicts the value for the response variable y as a straight line function of the value of the explanatory variable x. This line describes how a response variable y changes as an explanatory variable x changes.

Let $\hat{y}$ (y hat) denote the predicted value of y. The equation for the simple linear regression line is given by,

$$\hat{y} = b_0 + b_1 x$$

The constant term b_o denotes the **y-intercept** of the equation. The value of b_o is the height of the line above the value of $x = 0$.

The term b_1 denotes the ***slope*** of the regression equation. The value of b_1 is the amount by which y changes when x increases by ***one*** unit.

The prediction error or ***residual*** for an observation is the difference between the actual value and the predicted value of the response variable y.

Residual = $y - \hat{y}$ = *Observed – predicted.*

The formula for computing ***slope*** is, $b_1 = r\left(\frac{s_y}{s_x}\right)$

s_x is the standard deviation of the explanatory variable x,

s_y is the standard deviation of the response variable y. The value of r represents the correlation between x and y.

The formula for computing **y-intercept** is, $b_0 = \bar{y} - b\bar{x}$

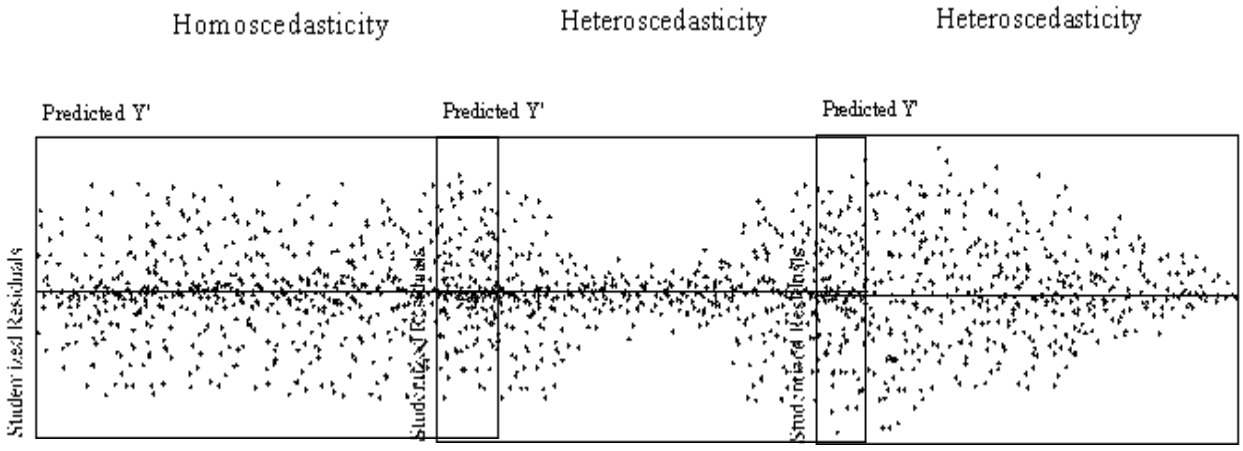
In the above equation, $\bar{y}$ is mean of y and $\bar{x}$ is the mean of x.

Proportional reduction in error (r^2) is the proportion of accuracy (in terms of explaining the variability) when regression equation is used to predict y values for each specific x values.

r-squared = r^2 = (correlation)^2.

The r^2 is also known as the **coefficient of determination**. An interpretation of r^2 =0.80 is: approximately 80% of the variability in the *response variable* can be explained by this linear regression equation.

Examples of random and non-random residual plots



Example: A random sample of 9 pair of data points representing home size and price are given below. Investigate the correlation between home size and price.

Home size (hundreds of square ft): 26, 27, 33, 29, 29, 34, 30, 40, 22
Home price (thousands of dollar): 259, 274, 294, 296, 325, 380, 457, 523, 215.

1. What are the intercept and slope of the model? Intercept: -140.194, slope: 15.89.

2. Write down the regression model. $\hat{y}$ = -140.194 + 15.89x.

3. What is the value of r^2? Give interpretation in the context of the problem.

   The r^2 value is 0.687. Approximately 69% of the variation in "House Price" can be explained by this regression model based on "House Size".

4. Predict the price of a house (in thousands of dollars) if the house size is 2700 sq. ft.

   $\hat{y}$ = -140.194 + 15.89(27) = 288.38. The predicted house price is $288,380.00.

5. If the true price for the house of 2700 sq. ft. in question 4 is $274,000, what is the model residual or prediction error?

   Model residual: y – $\hat{y}$ = 274 – 288.38 = – 14.38. The model is over estimating the house price by $14,380.