• **Regression Analysis** deals with the nature of the relationship between variables and

• **Correlation analysis** is concerned with measuring the strength between variables.
Regression and Correlation Analysis

- Regression
  - Simple Linear Regression
  - Multiple Regression

- Correlation Analysis
  - Simple Correlation Analysis
  - Multiple Correlation analysis
  - Partial Correlation Analysis
Regression and Correlation Analysis

• Develop an estimating equation
• Apply correlation analysis to determine the degree to which the variables are related
• Correlation analysis tells how well the estimating equation actually describes the relationship
Regression Analysis

- Develop a Scatter diagram to determine the relationship between variables

Positive correlation

Negative correlation
Regression Analysis

- Develop a Scatter diagram to determine the relationship between variables.

*Positive correlation*

*Negative correlation*
Scatter Diagram Method

Scatter Diagram is a graph of observed plotted points where each point represents the values of X and Y as a coordinate. It portrays the relationship between these two variables graphically.
Regression Analysis:

- Regression Analysis is concerned with the problem of describing or estimating the values of one variable, called dependent variable, on the basis of one or more other variables, called independent or explanatory variables.

- The objective of regression analysis is to arrive at an expression (‘model’) defining a line which is ‘best fits’ each set of plotted points. This line is called ‘Regression Line’.

- Equation of a straight line: \( Y = a + bx \)
  - \( Y \) is dependent variable & \( X \) is independent variable.
  - \( b \) is the slope (gradient) and \( a \) is the intercept.
Interpretation of the regression line

- $b$ measures the slope (gradient) of the line
- As $X$ changes, $y$ changes by $b$ times the change in $X$
- $a$ is the intercept
- When $X$ takes the value of zero, $Y = a$
- Given the equation for a straight line, it is possible to predict values of $Y$ for any given values of $X$
y = 1.2534x + 0.7888

Dependent variable Y
Independent variable X

Regression Line
The Regression Models

The population Regression Line

• The population Regression Line is a straight line relationship between $x$ and the mean of the $y$-values ($\mu_{y,x}$).

• The Population Regression line

$$\mu_{y,x} = \alpha + \beta x$$
The Population Regression Model

Population regression line: \( \mu_{y,x} = \alpha + \beta x \)

- Intercept: \( \alpha \)
- Slope: \( \beta \)
- \( \epsilon_i = \text{error} = Y_i - \mu_{y,x_i} \)
- \( Y_i = \text{observed value} \)
• Errors in the population model
\[ \epsilon_i = Y_i - \mu_{y.x_i} \quad \text{or} \quad Y_i = \mu_{y.x_i} + \epsilon_i \]
\[ \mu_{y.x_i} = Y_i - \epsilon_i \]

• Population Regression model:
\[ \mu_{y.x} = \alpha + \beta x \]
\[ Y_i - \epsilon_i = \alpha + \beta x_i \]
\[ Y_i = \alpha + \beta x_i + \epsilon_i \]
Deriving a best-fitting regression line – Least Squares Method

\[ \hat{y} = a + bx \]
Sample Regression Line: \( \hat{y} = a + bx \)

Residuals: \( e_i = y_i - \hat{y}_i \) \( \text{or} \)
\[
\begin{align*}
y_i &= \hat{y}_i + e_i
\end{align*}
\]

Sample Regression model: \( y_i = a + bx_i + e_i \)
Principle of the Least squares method

Method of least-squares estimations:

\[ e_i = y_i - \hat{y}_i \]

\( e_i \) referred to as “residuals” or “errors”

The sample regression line determined by minimizing \( \sum e_i^2 \) is called the least-squares regression line.

\[ \text{minimize } \sum e_i^2 = \text{minimize } \sum (y_i - \hat{y}_i)^2 \]

\( y_i \) - Observed value for dependent variable

\( \hat{y}_i \) — Computed value of the dependent variable for the \( i^{th} \) observation
Formula for Regression coefficients

\[ b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} \quad \text{or} \quad b = \frac{\sum xy - nxy}{\sum x^2 - n\bar{x}^2} \]

\[ a = \bar{y} - b\bar{x} \]

\( \Sigma x \) - The sum of the x
\( \Sigma y \) - The sum of the y
\( \Sigma x^2 \) - The sum of the squares of x
\( \Sigma xy \) - The sum of the products of x and y
The coefficients $a$ and $b$ can have negative as well as positive values.

A negative value for $b$ indicates an inverse relationship between $x$ and $y$, so that $y$ decreases as $x$ increases and vice versa.

A negative value for $a$ indicates a negative intercept on the $y$ axis.
Estimating population parameters

Sample Regression Line: \( \hat{y} = a + bx \)
The Population Regression line: \( \mu_{y|x} = \alpha + \beta x \)

• The sample value of \( a \) is the best estimator of \( \alpha \), while the sample value of \( b \) is the best estimator of \( \beta \)

• Values of \( a \) and \( b \) together with a given value of \( x \) yield a predicted value of \( y \), which is denoted \( \hat{y} \), \( \hat{y} \) is the best estimator of the population value \( \mu_{y|x} \)

• \( e_i \) is the best estimator of the population value \( \varepsilon_i \)
The vice president for research and development of a chemical and fiber manufacturing company believes that the firm’s annual profits depend on the amount spent on R&D. The new chief executive officer does not agree and has asked for evidence. Here are data for 6 years:

<table>
<thead>
<tr>
<th>Year</th>
<th>Millions Spent on R&amp;D (x)</th>
<th>Annual Profit (millions) (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>2015</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>2016</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>2017</td>
<td>11</td>
<td>40</td>
</tr>
<tr>
<td>2018</td>
<td>5</td>
<td>31</td>
</tr>
</tbody>
</table>

The vice president for R&D wants an equation for predicting annual profits from the amount budgeted for R&D.
Scatter Diagram

Annual profit (mn) vs. Amount spent on R&D (mn)
1. Compute best-fitting regression equation between annual profits \( Y \) and amount budgeted for R&D \( x \).

\[
b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}
\]

\[
a = \bar{y} - b \bar{x}
\]
<table>
<thead>
<tr>
<th>Year</th>
<th>x</th>
<th>Y</th>
<th>$x^2$</th>
<th>$xy$</th>
<th>$y^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2</td>
<td>20</td>
<td>4</td>
<td>40</td>
<td>400</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
<td>25</td>
<td>9</td>
<td>75</td>
<td>625</td>
</tr>
<tr>
<td>2015</td>
<td>5</td>
<td>34</td>
<td>25</td>
<td>170</td>
<td>1156</td>
</tr>
<tr>
<td>2016</td>
<td>4</td>
<td>30</td>
<td>16</td>
<td>120</td>
<td>900</td>
</tr>
<tr>
<td>2017</td>
<td>11</td>
<td>40</td>
<td>121</td>
<td>440</td>
<td>1600</td>
</tr>
<tr>
<td>2018</td>
<td>5</td>
<td>31</td>
<td>25</td>
<td>155</td>
<td>961</td>
</tr>
</tbody>
</table>

\[
\sum x = 30 \quad \sum y = 180 \quad \sum x^2 = 200 \quad \sum xy = 1000 \quad \sum y^2 = 5642
\]

\[
b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{6(1000) - (30 \times 180)}{6(200) - 900} = \frac{6000 - 5400}{1200 - 900} = \frac{600}{300} = 2
\]

\[
a = \bar{y} - b\bar{x} = 30 - 2 \times 5 = 30 - 10 = 20
\]

\[
a = 20
\]

Regression equation: $Y = 20 + 2x$
Scattering of points around the regression line

$y = 2x + 20$
2. What would the regression model estimate for annual profits be when the amount budgeted for R&D is 7 millions?

\[ \hat{y} = 20 + 2x \]
\[ = 20 + 2 \times 7 \]
\[ = 34 \text{ i.e. 34 millions.} \]

\( b = 2 \) means profits rise by 2 million for a unit increase (1 million) in amount spent on R&D.
Measures of Goodness of fit in regression analysis

1. Standard error of estimate ($S_e$)
   • Measures the reliability (accuracy) of the estimating equation.
   • $S_e$ is one of the most useful measures of goodness of fit in regression analysis.
   • Standard error of estimate ($S_e$) measures the variability or scatter of the observed values around the regression line.
   • This sample statistics ($S_e$) is the standard deviation of the errors ($e_i$) about the sample regression line.
\[ S_e = \sqrt{\frac{\Sigma (y - \hat{y})^2}{n-2}} \quad \text{or} \quad S_e = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b \Sigma xy}{n - 2}} \quad \sum (y - \hat{y})^2 = \sum e_i^2 \]

\[ S_e = \sqrt{\frac{\Sigma y^2 - a \Sigma y - b \Sigma xy}{n - 2}} \]

\[ S_e = \sqrt{\frac{5642 - 20(180) - 2(1000)}{6 - 2}} \]

\[ = 10.5 \]
Interpreting the Standard error of estimate ($S_e$)

- The larger the standard error of estimate, the greater the scattering (or dispersion) of points around the regression line.

- If $S_e = 0$, the estimating equation to be a “Perfect” estimator of the dependent variable. In that case all the data points would lie directly on the regression line and no points would be scattered around it.
Measures of Goodness of fit in regression analysis

2. Coefficient of determination ($r^2$)

- The Coefficient of determination can be used to measure the extent, or strength, of the association that exists between two variables.
- $r^2$ is a measure of the degree of linear association between $x$ and $y$.

Variation of the $Y$ values around the regression line $= \sum (y - \hat{y})^2$

Variation of the $Y$ values around their own mean $= \sum (y - \bar{y})^2$

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

$$r^2 = \frac{a \sum y + b \sum xy - n\bar{y}^2}{\sum y^2 - n\bar{y}^2}$$
<table>
<thead>
<tr>
<th>Year</th>
<th>$x$</th>
<th>$y$</th>
<th>$\hat{y}$</th>
<th>$(y - \hat{y})^2$</th>
<th>$(y - \bar{y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>2</td>
<td>20</td>
<td>24</td>
<td>16</td>
<td>100</td>
</tr>
<tr>
<td>2014</td>
<td>3</td>
<td>25</td>
<td>26</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>2015</td>
<td>5</td>
<td>34</td>
<td>30</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>2016</td>
<td>4</td>
<td>30</td>
<td>28</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2017</td>
<td>11</td>
<td>40</td>
<td>42</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>2018</td>
<td>5</td>
<td>31</td>
<td>30</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum x = 30$</td>
<td>$\sum y = 180$</td>
<td>$\sum (y - \hat{y})^2 = 42$</td>
<td>$\sum (y - \bar{y})^2 = 242$</td>
<td></td>
</tr>
</tbody>
</table>

$\bar{y} = 30$  
*Regression equation: $\hat{y} = 20 + 2x$*

$$r^2 = 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} = 1 - \frac{42}{242} = 0.826$$
Interpretation of $r^2$

- The amount of the variation in Y that is explained by the regression line.
- (eg: 82.6 of the variation in Y is explained by the regression line)
- $r^2$ lies somewhere between 1 and 0.
- $r^2$ close to 1 indicates a strong correlation between x and y.
- $r^2$ near 0 means that there is little correlation between x and y.
- The value of $r^2$ is zero when there is no correlation.