Chapter 4

Curve Fitting & Correlation

4.1 Introduction
The process of constructing an approximate curve \( y = f(x) \), which fit best to a given discrete set of points \((x_i, y_i), \ i = 1, 2, 3, ..., n\) is called curve fitting. Curve fitting and interpolation are closely associated procedures. In interpolation, the fitted function should pass through all given data points; whereas curve fitting methodologically fits a unique curve to the data points, which may or may not lie on the fitted curve. The difference between interpolation and curve fitting; while attempting to fit a linear function; is illustrated in the adjoining figure.

4.2 Principle of Least Squares
The principle of least squares is one of the most popular methods for finding the curve of best fit to a given data set \((x_i, y_i), \ i = 1, 2, 3, ..., n\).

Let \( y = f(x) \) be the equation of the curve to be fitted to the given set of points \( P_1(x_1, y_1), P_2(x_2, y_2), \ldots, P_n(x_n, y_n) \).

Then at a point \( x = x_i \), the observed value of the ordinate is \( P_iM_i = y_i \) say and let the expected (theoretical) value be \( f(x_i) \), shown by \( L_iM_i \) in the adjoining figure. The difference between the observed and expected values is the error \( e_i = P_iL_i \).

Then \( e_1 = y_1 - f(x_1) \)
\[ e_2 = y_2 - f(x_2) \]
\[ \vdots \]
\[ e_n = y_n - f(x_n) \]

Squaring each error \( e_i \) (to take care of negative errors) and adding, we get
\[ E = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2 \]. The curve of best fit is that for which \( E \) is minimum. This is called the Principle of least squares.
4.2.1 Fitting a straight line \( y = ax + b \).

Let \( y = ax + b \) be the straight line to be fitted to the given set of data points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).

Then \( e_i = y_i - f(x_i) = y_i - (ax_i + b), \quad i = 1, 2, \ldots, n \)

\[ E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \]

Now by the principle of least square, for the curve of best fit, \( E \) is minimum

\[ \frac{\partial E}{\partial a} = 0 \quad \text{and} \quad \frac{\partial E}{\partial b} = 0 \]

\[ \Rightarrow \sum_{i=1}^{n} 2 \left(y_i - (ax_i + b)\right)(-x_i) = 0 \quad \Rightarrow \sum_{i=1}^{n} 2 \left(y_i - ax_i - b\right)(-1) = 0 \]

\[ \Rightarrow \sum_{i=1}^{n} (y_i x_i - ax_i^2 - bx_i) = 0 \quad \Rightarrow \sum_{i=1}^{n} (y_i - ax_i - b) = 0 \]

\[ \Rightarrow \sum_{i=1}^{n} y_i x_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i \quad \Rightarrow \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \]

\[
\begin{align*}
\text{①} & & \text{②}
\end{align*}
\]

Since \( x_i, y_i \) are known set of values, equation ① and ② are two equation with variables \( a \) and \( b \), ignoring the suffices, ② and ① can be rewritten as

\[ \sum y = a \sum x + nb \quad \ldots \text{③} \]

and \[ \sum xy = a \sum x^2 + b \sum x \quad \ldots \text{④} \]

③and ④are known as Normal equation for fitting a straight line \( y = ax + b \)

➢ If the equation of line is taken as \( y = a + bx \)

we get normal equations as: \( \sum y = na + b \sum x \) and \( \sum xy = a \sum x + b \sum x^2 \)

**Example 1** By the method of least squares, find a straight line that best fits the following data points.

<table>
<thead>
<tr>
<th>x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>1.0</td>
<td>2.9</td>
<td>4.8</td>
<td>6.7</td>
<td>8.6</td>
</tr>
</tbody>
</table>

**Solution:** Let line of best fit be given by \( y = ax + b \) \ldots ①

Normal equations are given by:

\[ \sum y = a \sum x + nb \quad \ldots \text{②} \]

and \[ \sum xy = a \sum x^2 + b \sum x \quad \ldots \text{③} \]

Calculating \( \sum x, \sum y, \sum xy \) and \( \sum x^2 \)
Substituting values of $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$ in (2) and (3)

\[24 = 10a + 5b \quad \ldots \text{④}\]

and

\[67 = 30a + 10b \quad \ldots \text{⑤}\]

Solving ④ and ⑤, we get $a = 1.9$ and $b = 1$

Substituting in ①, line of best fit is $y = 1.9x + 1$

**Example 2** Fit a straight line to following data

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>1.0</td>
<td>1.8</td>
<td>3.3</td>
<td>4.5</td>
<td>6.3</td>
</tr>
</tbody>
</table>

**Solution:** Let line of best fit be given by $y = ax + b \quad \ldots \text{①}$

Normal equations are given by:

\[\sum y = a \sum x + nb \quad \ldots \text{②}\]

and

\[\sum xy = a \sum x^2 + b \sum x \quad \ldots \text{③}\]

Calculating $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$xy$</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1.8</td>
<td>1.8</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3.3</td>
<td>6.6</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4.5</td>
<td>13.5</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>6.3</td>
<td>25.2</td>
<td>16</td>
</tr>
</tbody>
</table>

$\sum x = 10 \quad \sum y = 16.9 \quad \sum xy = 47.1 \quad \sum x^2 = 30$

Substituting values of $\sum x$, $\sum y$, $\sum xy$ and $\sum x^2$ in ② and ③

\[16.9 = 10a + 5b \quad \ldots \text{④}\]
and \[ 47.1 = 30a + 10b \] ... ⑤

Solving ④ and ⑤, we get \[ a = \frac{133}{100} = 1.33 \] and \[ b = \frac{18}{25} = 0.72 \]

Substituting in ①, line of best fit is \[ y = 1.33x + 0.72 \]

**Example 3** If \( F \) is the force required to lift a load \( W \), by means of a pulley, fit a linear expression \( F = a + bW \) against the following data:

<table>
<thead>
<tr>
<th>( W )</th>
<th>50</th>
<th>70</th>
<th>100</th>
<th>120</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td>12</td>
<td>15</td>
<td>21</td>
<td>25</td>
</tr>
</tbody>
</table>

**Solution:** Line for best fit is given as \[ F = a + bW \] ... ①

Normal equations are given by:

\[ \sum F = na + b \sum W \] ... ②

and

\[ \sum WF = a \sum W + b \sum W^2 \] ... ③

<table>
<thead>
<tr>
<th>( W )</th>
<th>( F )</th>
<th>( WF )</th>
<th>( W^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>12</td>
<td>600</td>
<td>2500</td>
</tr>
<tr>
<td>70</td>
<td>15</td>
<td>1050</td>
<td>4900</td>
</tr>
<tr>
<td>100</td>
<td>21</td>
<td>2100</td>
<td>10000</td>
</tr>
<tr>
<td>120</td>
<td>25</td>
<td>3000</td>
<td>14400</td>
</tr>
</tbody>
</table>

\[ \sum W = 340 \] \[ \sum F = 73 \] \[ \sum WF = 6750 \] \[ \sum W^2 = 31800 \]

Substituting values of \( \sum W, \sum F, \sum WF \) and \( \sum W^2 \) in ② and ③

\[ 73 = 4a + 340b \] ... ④

and

\[ 6750 = 340a + 31800b \] ... ⑤

Solving ④ and ⑤, we get \( a = 2.2759 \) and \( b = 0.1879 \)

Substituting in ①, line of best fit is \( F = 2.2759 + 0.1879W \)

**4.2.2 Fitting a parabola \( y = ax^2 + bx + c \)**

Let \( y = ax^2 + bx + c \) be the parabola to be fitted to the given set of data points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\).

Then \( e_i = y_i - f(x_i) = y_i - (ax_i^2 + bx_i + c) \), \( i = 1, 2, \ldots, n \)

\[ E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (ax_i^2 + bx_i + c))^2 \]

Now by the principle of least square, for the curve of best fit, \( E \) is minimum

\[ : \frac{\partial E}{\partial a} = 0 \text{, } \frac{\partial E}{\partial b} = 0 \text{ and } \frac{\partial E}{\partial c} = 0 \]
Solving we get normal equations as:

\[ \sum y = a \sum x^2 + b \sum x + nc \]
\[ \sum xy = a \sum x^3 + b \sum x^2 + c \sum x \]
\[ \sum x^2y = a \sum x^4 + b \sum x^3 + c \sum x^2 \]

**Example 4** Fit a parabola \( y = ax^2 + bx + c \) to the given data

| \( x \) | 10 | 12 | 15 | 23 | 20 |
| \( y \) | 14 | 17 | 23 | 25 | 21 |

**Solution:** Let the parabola of best fit be given by \( y = ax^2 + bx + c \) \(... 1\)

Normal equations are given by:

\[ \sum y = a \sum x^2 + b \sum x + nc \] \(... 2\)
\[ \sum xy = a \sum x^3 + b \sum x^2 + c \sum x \] \(... 3\)
\[ \sum x^2y = a \sum x^4 + b \sum x^3 + c \sum x^2 \] \(... 4\)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( xy )</th>
<th>( x^2 )</th>
<th>( x^2y )</th>
<th>( x^3 )</th>
<th>( x^4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>14</td>
<td>140</td>
<td>100</td>
<td>1400</td>
<td>1000</td>
<td>10000</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>204</td>
<td>144</td>
<td>2448</td>
<td>1728</td>
<td>20736</td>
</tr>
<tr>
<td>15</td>
<td>23</td>
<td>345</td>
<td>225</td>
<td>5175</td>
<td>3375</td>
<td>50625</td>
</tr>
<tr>
<td>23</td>
<td>25</td>
<td>575</td>
<td>529</td>
<td>13225</td>
<td>12167</td>
<td>279841</td>
</tr>
<tr>
<td>20</td>
<td>21</td>
<td>420</td>
<td>400</td>
<td>8400</td>
<td>8000</td>
<td>160000</td>
</tr>
<tr>
<td>( \sum x = \sum y = \sum xy = \sum x^2 = \sum x^2y = \sum x^3 = \sum x^4 = ) 80</td>
<td>100</td>
<td>1684</td>
<td>1398</td>
<td>30684</td>
<td>26270</td>
<td>521202</td>
</tr>
</tbody>
</table>

Substituting values of \( \sum x, \sum y, \sum xy \) and \( \sum x^2 \) in \( 2 \) and \( 3 \) and \( 4 \)

\[ 100 = 1398a + 80b + 5c \] \(... 5\)
\[ 1684 = 26270a + 1398b + 80c \] \(... 6\)
\[ 30648 = 521202a + 26270b + 1398c \] \(... 7\)

Solving \( 5 \), \( 6 \) and \( 7 \), we get \( a = -0.07 \), \( b = 3.01 \), \( c = -8.73 \)

Substituting in \( 1 \), parabola of best fit is \( y = -0.07 x^2 + 3.01x - 8.73 \)

**Example 5** Fit a 2\(^{nd}\) parabola to the given data

| \( x \) | 1  | 3  | 4  | 6  | 8  | 9  | 11 | 14 |
| \( y \) | 1  | 2  | 4  | 4  | 5  | 7  | 8  | 9  |

**Solution:** Let the parabola of best fit be given by \( y = ax^2 + bx + c \) \(... 1\)
Normal equations are given by:

\[ \sum y = a \sum x^2 + b \sum x + nc \]  \hspace{1cm} \ldots \text{(2)}

\[ \sum xy = a \sum x^3 + b \sum x^2 + c \sum x \]  \hspace{1cm} \ldots \text{(3)}

\[ \sum x^2y = a \sum x^4 + b \sum x^3 + c \sum x^2 \]  \hspace{1cm} \ldots \text{(4)}

Substituting values of \( \sum x, \sum y, \sum xy \) and \( \sum x^2 \) in (2) and (3) and (4)

\[ 40 = 524a + 56b + 8c \]  \hspace{1cm} \ldots \text{(5)}

\[ 364 = 5624a + 524b + 56c \]  \hspace{1cm} \ldots \text{(6)}

\[ 3846 = 65348a + 5624b + 524c \]  \hspace{1cm} \ldots \text{(7)}

Solving (5) (6) and (7), we get

\[ a = \frac{103}{11229} = 0.009, \quad b = \frac{8672}{11229} = 0.77, \quad c = \frac{4375}{22458} = 0.195 \]

Substituting in (1), parabola of best fit is \( y = 0.009x^2 + 0.77x + 0.195 \)

### 4.2.3 Change of Scale
If the data values are equispaced (with height \( h \)) and quite large for computation, simplification may be done by origin shifting as given below:

- When number of observations \( n \) is odd, take the origin at middle value of the table; say \( x_0 \) and substitute \( u = \frac{x-x_0}{h} \)

- \( y \) values if small; may be left unchanged; or we can shift them at average value of \( y \) data \( v = \frac{y-y_0}{h} \)
- When number of observations \((n)\) is even, take the origin as mean of two middle values, with new height \(\frac{h}{2}\) and substitute \(u = \frac{x-x_0}{h/2}\)

**Example 6**  Fit a 2\(^{nd}\) degree parabola for the following data:

<table>
<thead>
<tr>
<th>(x)</th>
<th>1929</th>
<th>1930</th>
<th>1931</th>
<th>1932</th>
<th>1933</th>
<th>1934</th>
<th>1935</th>
<th>1936</th>
<th>1937</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>352</td>
<td>356</td>
<td>357</td>
<td>358</td>
<td>360</td>
<td>361</td>
<td>361</td>
<td>360</td>
<td>359</td>
</tr>
</tbody>
</table>

**Solution:** Since number of observations is odd and \(h = 1\),

taking \(x_0 = 1933, y_0 = 357, u = x - 1933, v = y - 357\)

The equation \(y = ax^2 + bx + c\) is transformed to \(v = Au^2 + Bu + C\)

Normal equations are

\[
\sum v = A \Sigma u^2 + B \Sigma u + 9c .... \(2\)
\]
\[
\sum uv = A \Sigma u^3 + B \Sigma u^2 + c \Sigma u .... \(3\)
\]
\[
\sum u^2v = A \Sigma u^4 + B \Sigma u^3 + c \Sigma u^2 .... \(4\)
\]

Calculating \(\sum u, \sum u^2, \sum u^3, \sum u^4, \sum v, \sum uv\) and \(\sum u^2v\)

<table>
<thead>
<tr>
<th>(x)</th>
<th>(u)</th>
<th>(y)</th>
<th>(v)</th>
<th>(uv)</th>
<th>(u^2)</th>
<th>(u^2v)</th>
<th>(u^3)</th>
<th>(u^4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1929</td>
<td>-4</td>
<td>352</td>
<td>-5</td>
<td>20</td>
<td>16</td>
<td>-80</td>
<td>-64</td>
<td>256</td>
</tr>
<tr>
<td>1930</td>
<td>-3</td>
<td>356</td>
<td>-1</td>
<td>3</td>
<td>9</td>
<td>-9</td>
<td>-27</td>
<td>81</td>
</tr>
<tr>
<td>1931</td>
<td>-2</td>
<td>357</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>-8</td>
<td>16</td>
</tr>
<tr>
<td>1932</td>
<td>-1</td>
<td>358</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1933</td>
<td>0</td>
<td>360</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1934</td>
<td>1</td>
<td>361</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1935</td>
<td>2</td>
<td>361</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>16</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>1936</td>
<td>3</td>
<td>360</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>27</td>
<td>27</td>
<td>81</td>
</tr>
<tr>
<td>1937</td>
<td>4</td>
<td>359</td>
<td>2</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>256</td>
</tr>
</tbody>
</table>

\[
\sum u = 0, \quad \sum v = 11, \quad \sum uv = 51, \quad \sum u^2 = 60, \quad \sum u^2v = -9, \quad \sum u^3 = 0, \quad \sum u^4 = 708
\]

Substituting \(\sum u, \sum u^2, \sum u^3, \sum u^4, \sum v, \sum uv\) and \(\sum u^2v\) in \(2\) and \(3\) and \(4\)

\[
\Rightarrow 11 = 60A + 9C .... \(5\)
\]
\[
51 = 60B .... \(6\)
\]
\[
-9 = 708A + 60C .... \(7\)
\]

Solving \(5\), \(6\) and \(7\), we get \(A = \frac{-247}{924}\) and \(B = \frac{17}{20}\), \(C = \frac{694}{231}\)
Substituting in (1), parabola of best fit is
\[ v = \frac{-247}{924} x^2 + \frac{17}{20} x + \frac{694}{231} \]
\[ \Rightarrow y - 357 = \frac{-247}{924} (x - 1933)^2 + \frac{17}{20} (x - 1933) + \frac{694}{231} \]
\[ \Rightarrow y = -0.267x^2 + 1034.29x - 1000106.41 \]

**Example 7** The weight of a calf taken at end of every month is given below. Fit a straight line using the method of least squares. Also compute monthly growth rate.

<table>
<thead>
<tr>
<th>x</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>52.5</td>
<td>58.7</td>
<td>65.0</td>
<td>70.2</td>
<td>75.4</td>
<td>81.1</td>
<td>87.2</td>
<td>95.5</td>
<td>102.2</td>
<td>108.4</td>
</tr>
</tbody>
</table>

**Solution:** Here \( n = 10 \) is even, \( \therefore \) taking origin at \( \frac{5+6}{2} = 5.5 \) and new height as

\[ \frac{h}{2} = 0.5 \quad \therefore u = \frac{x-5.5}{0.5} \text{ and let } v = y \]

Let line of best fit \( y = ax + b \) be transformed to \( v = Au + B \) \( \ldots(1) \)

Normal equations are given by
\[ \Sigma v = a \Sigma u + nb \quad \ldots(2) \]

and
\[ \Sigma uv = a \Sigma u^2 + b \Sigma u \quad \ldots(3) \]

<table>
<thead>
<tr>
<th>x</th>
<th>( u = \frac{x - 5.5}{0.5} )</th>
<th>( v = y )</th>
<th>( uv )</th>
<th>( u^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-9</td>
<td>52.5</td>
<td>-472.5</td>
<td>81</td>
</tr>
<tr>
<td>2</td>
<td>-7</td>
<td>58.7</td>
<td>-410.9</td>
<td>49</td>
</tr>
<tr>
<td>3</td>
<td>-5</td>
<td>65.0</td>
<td>-325.0</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>-3</td>
<td>70.2</td>
<td>-210.6</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>-1</td>
<td>75.4</td>
<td>-75.4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>81.1</td>
<td>81.1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>87.2</td>
<td>261.6</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>95.5</td>
<td>477.5</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>102.2</td>
<td>715.4</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>108.4</td>
<td>975.6</td>
<td>81</td>
</tr>
</tbody>
</table>

\( \Sigma u = 0 \)
\( \Sigma v = 796.2 \)
\( \Sigma uv = 1016.8 \)
\( \Sigma u^2 = 330 \)

Substituting values of \( \Sigma u, \Sigma v, \Sigma uv \) and \( \Sigma u^2 \) in (2) and (3)
\[ 796.2 = 10B \quad \text{and} \quad 1016.8 = 330a \]

\( \therefore A = 3.081 \text{ and } B = 79.62 \)

Substituting in (1), line of best fit is \( v = 3.081u + 79.62 \)
\[ y = 3.081 \left( \frac{x - 5.5}{0.5} \right) + 79.62 \quad \therefore \text{Line of best fit is } y = 6.162x + 45.729 \]

Average growth rate per month is given by: \( \frac{dy}{dx} = 6.162 \)

4.3 Correlation

Correlation is a measure of association between two variables; which may be dependent or independent. Whenever two variables \( x \) and \( y \) are so related; that increase in one in accompanied by an increase or decrease in the other, then the variables are said to be correlated. Coefficient of correlation \( (r) \) lies between \(-1\) and \(+1\), i.e. \(-1 \leq r \leq 1\).

If \( r \) is zero; no correlation between two variables, positive correlation \( (0 < r \leq +1) \); when both variables increase or decrease simultaneously, and negative correlation \( (-1 \leq r < 0) \); when increase in one is associated with decrease in other variable and vice-versa.

4.4 Karl Pearson Coefficient of Correlation

Coefficient of correlation \( (r) \) between two variables \( x \) and \( y \) is defined as

\[
r = \frac{\text{Covariance (x,y)}}{\sqrt{\text{Variance (x)}} \sqrt{\text{Variance (y)}}} = \frac{\Sigma d_x d_y}{\sqrt{(\Sigma d_x^2)(\Sigma d_y^2)}} = \frac{\rho}{\sigma_x \sigma_y}
\]

where \( d_x = x - \bar{x}, \quad d_y = y - \bar{y}, \quad \bar{x}, \bar{y} \) are means of \( x \) and \( y \) data values.

\( \rho = \text{Cov} (x, y) = \frac{\Sigma d_x d_y}{n} \) is the covariance between the variables \( x \) and \( y \).

Also \( \sigma_x = \sqrt{\frac{\Sigma d_x^2}{n}} \) and \( \sigma_y = \sqrt{\frac{\Sigma d_y^2}{n}} \)

Example 8 If \( \text{Cov} (x, y) = 10, \text{var}(x) = 25, \text{var}(y) = 9 \) find coefficient of correlation.

Solution: \( r = \frac{\text{Cov} (x, y)}{\sqrt{\text{Var}(x)} \sqrt{\text{Var}(y)}} = \frac{10}{\sqrt{25} \sqrt{9}} = \frac{10}{5 \times 3} = 0.67 \)

Example 9 Calculate coefficient of correlation from the following data:

<table>
<thead>
<tr>
<th>( x )</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>15</td>
<td>16</td>
<td>14</td>
<td>13</td>
<td>11</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Solution: Karl Pearson coefficient of correlation \( (r) \) is given by: \( r = \frac{\Sigma d_x d_y}{\sqrt{(\Sigma d_x^2)(\Sigma d_y^2)}} \)

where \( d_x = x - \bar{x}, \quad d_y = y - \bar{y}, \quad \bar{x}, \bar{y} \) are means of \( x \) and \( y \) data values.
Shortcut Method for Karl Pearson Coefficient of Correlation

We can also find Karl Pearson Coefficient of Correlation by taking assumed means as shown:

If we take \( d_x = x - a \), \( d_y = y - b \),

where \( a \) and \( b \) are assumed means of \( x \) and \( y \) data values

Then \[
r = \frac{\Sigma d_x d_y - \frac{1}{n} \Sigma d_x \Sigma d_y}{\sqrt{\left(\Sigma d_x^2\right) - \frac{1}{n} \left(\Sigma d_x\right)^2} \sqrt{\left(\Sigma d_y^2\right) - \frac{1}{n} \left(\Sigma d_y\right)^2}}
\]

\( \therefore \) If \( x_i \)'s are equispaced with height \( h \), we can take \( d_x = \frac{x - a}{h} \)

Similarly if \( y_i \)'s are equispaced with height \( k \), then \( d_y = \frac{y - b}{k} \)

**Example 10** Calculate coefficient of correlation from the following data:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( d_x = (x - \bar{x}) )</th>
<th>( d_x^2 )</th>
<th>( y )</th>
<th>( d_y = (y - \bar{y}) )</th>
<th>( d_y^2 )</th>
<th>( d_x d_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>4</td>
<td>16</td>
<td>15</td>
<td>3</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>9</td>
<td>16</td>
<td>4</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>4</td>
<td>14</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-1</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-2</td>
<td>2</td>
<td>10</td>
<td>-2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>-3</td>
<td>9</td>
<td>8</td>
<td>-4</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>-4</td>
<td>16</td>
<td>9</td>
<td>-3</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

\( \Sigma x = 45 \)
\( \bar{x} = 5 \)
\( \Sigma d_x^2 = 60 \)
\( \Sigma y = 108 \)
\( \bar{y} = 12 \)
\( \Sigma d_y^2 = 60 \)
\( \Sigma d_x d_y = 57 \)

\[
r = \frac{\Sigma d_x d_y - \frac{1}{n} \Sigma d_x \Sigma d_y}{\sqrt{\left(\Sigma d_x^2\right) - \frac{1}{n} \left(\Sigma d_x\right)^2} \sqrt{\left(\Sigma d_y^2\right) - \frac{1}{n} \left(\Sigma d_y\right)^2}}
\]

\[
r = \frac{57}{\sqrt{60 \times 60}} = \frac{57}{60} = 0.95
\]
4.4.2 Coefficient of Correlation of Bivariate Frequency Distribution

If given data is in the form of a bivariate frequency distribution, then

\[
\text{Then } r = \frac{\sum f dx dy - \frac{1}{n} \sum f d_x \sum f d_y}{\sqrt{\left(\sum f d_x^2\right) - \frac{1}{n} \left(\sum f d_x\right)^2} \sqrt{\left(\sum f d_y^2\right) - \frac{1}{n} \left(\sum f d_y\right)^2}} , \ n = \sum f
\]

Example 11 Following table gives a bivariate distribution showing frequency of marks obtained according to age by a group of 52 students in an intelligent test:

<table>
<thead>
<tr>
<th>Age</th>
<th>Marks</th>
<th>16-18</th>
<th>18-20</th>
<th>20-22</th>
<th>22-24</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-30</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>30-40</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>40-50</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>50-60</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>60-70</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compute the correlation between marks and age of the students.

Solution: Let marks obtained by the students be denoted by \( x \) and age by \( y \), then coefficient of correlation \( (r) \) for the bivariate frequency distribution is given by:

\[
r = \frac{\sum f dx dy - \frac{1}{n} \sum f d_x \sum f d_y}{\sqrt{\left(\sum f d_x^2\right) - \frac{1}{n} \left(\sum f d_x\right)^2} \sqrt{\left(\sum f d_y^2\right) - \frac{1}{n} \left(\sum f d_y\right)^2}} , \ n = \sum f
\]
Let \( d_x = \frac{x-a}{10}, \quad d_y = \frac{y-b}{2} \), where \( a \) & \( b \) denote assumed mean classes

Here \( a \) is taken as \( 30 - 40 \), \( b \) is taken as \( 18 - 20 \)

Also quantities in brackets denote \( d_x d_y \) for each cell.

\[ : f_d x d_y \text{ for each cell is obtained by multiplying frequency of each cell with } d_x d_y \text{ and added across rows or columns to get } f_d x d_y \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>16-18</th>
<th>18-20</th>
<th>20-24</th>
<th>22-24</th>
<th>( F )</th>
<th>( d_x )</th>
<th>( f_d x )</th>
<th>( f_d x^2 )</th>
<th>( f_d x d_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-20</td>
<td>2 (2)</td>
<td>1 (0)</td>
<td>1 (-2)</td>
<td>-</td>
<td>4</td>
<td>-2</td>
<td>-8</td>
<td>16</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>20-30</td>
<td>3 (1)</td>
<td>2 (0)</td>
<td>3 (-1)</td>
<td>2 (-2)</td>
<td>10</td>
<td>-1</td>
<td>-10</td>
<td>10</td>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>30-40</td>
<td>3 (0)</td>
<td>4 (0)</td>
<td>5 (0)</td>
<td>6 (0)</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>40-50</td>
<td>2 (-1)</td>
<td>2 (0)</td>
<td>3 (1)</td>
<td>4 (2)</td>
<td>11</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>50-60</td>
<td>-</td>
<td>1 (0)</td>
<td>2 (2)</td>
<td>2 (4)</td>
<td>5</td>
<td>2</td>
<td>10</td>
<td>20</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>60-70</td>
<td>-</td>
<td>1 (0)</td>
<td>2 (3)</td>
<td>1 (6)</td>
<td>4</td>
<td>3</td>
<td>12</td>
<td>36</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>( F )</td>
<td>10</td>
<td>11</td>
<td>16</td>
<td>15</td>
<td>52</td>
<td>Totals</td>
<td>15</td>
<td>93</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>

\[ d_y \] -1 | 0 | 1 | 2 | Totals
\[ f_d y \] -10 | 0 | 16 | 30 | 36
\[ f_d y^2 \] 10 | 0 | 16 | 60 | 86
\[ f_d x d_y \] 5 | 0 | 8 | 18 | 31

\[ \therefore \rho = \frac{31 - \frac{1}{52}(15)(36)}{\sqrt{\frac{(93)(31)}{52}} \sqrt{\frac{1}{52}(86)} = \frac{20.6154}{9.4166(7.8152)} = 0.2801 \]

Thus there is a weak positive correlation between marks and age of the students.

**4.4.3 Coefficient of Correlation by Rank differences**

Rank correlation is used for attributes (like beauty, intelligence etc.) which cannot be measured quantitatively but can be provided with comparative ranks.

**Spearman’s Rank Correlation** in given by: \( r = 1 - \frac{6 \Sigma D^2}{n(n^2-1)} \), where \( D = R_1 - R_2 \)

**Tied Ranks:** If two or more observations in a data are equal, each observation is provided with an average rank and a correction factor is applied to correlation formula given as: Correction Factor (C.F.) = \( \sum m(m^2 - 1) \), \( m \) is the number of times each observation is repeated.
Spearman’s Rank Correlation for repeated ranks is given by:

\[ r = 1 - \frac{6\left(\sum D^2 + \frac{1}{12}\text{C.F.}\right)}{n(n^2-1)}, \text{ where } D = R_1 - R_2 \]

**Example 12** Calculate the coefficient of correlation from the following data; given ranks of 10 students in English and Mathematics.

<table>
<thead>
<tr>
<th>Rank in English</th>
<th>3</th>
<th>1</th>
<th>5</th>
<th>4</th>
<th>2</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>9</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank in Mathematics</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
</tbody>
</table>

**Solution:** Since comparative ranks are given; instead of marks, using Spearman’s Rank Correlation is given by:

\[ r = 1 - \frac{6\sum D^2}{n(n^2-1)}, \text{ where } D = R_1 - R_2 \]

<table>
<thead>
<tr>
<th>Rank in English</th>
<th>Rank in Mathematics</th>
<th>( D = R_1 - R_2 )</th>
<th>( D^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

\[ \sum D^2 = 52 \]

\[ \therefore r = 1 - \frac{6(52)}{10(10^2-1)} = 0.6848 \]

**Example 13** Eight competitors in a beauty contest got marks (out of 10) by three judges as given:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>9</th>
<th>6</th>
<th>5</th>
<th>10</th>
<th>3</th>
<th>1</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge B</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Judge C</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

Use rank correlation to discuss which pair of judges has the nearest approach to common tastes in beauty.
Solution: Since instead of ranks, marks are given by the three judges, converting the given data to comparative ranks for the eight competitors

<table>
<thead>
<tr>
<th>Judge A</th>
<th>Judge B</th>
<th>Judge C</th>
<th>$D_{AB}$</th>
<th>$D_{AB}^2$</th>
<th>$D_{BC}$</th>
<th>$D_{BC}^2$</th>
<th>$D_{AC}$</th>
<th>$D_{AC}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks</td>
<td>Rank</td>
<td>Marks</td>
<td>Rank</td>
<td>Marks</td>
<td>Rank</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>-4</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>-4</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>9</td>
<td>-5</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>49</td>
<td>-6</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>-2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

Here $D_{AB} = \text{Rank by Judge A} - \text{Rank by Judge B}$, also $\sum D_{AB}^2 = 100$

Similarly $D_{BC} = \text{Rank by Judge B} - \text{Rank by Judge C}$, also $\sum D_{BC}^2 = 120$

$D_{AC} = \text{Rank by Judge A} - \text{Rank by Judge C}$, also $\sum D_{AC}^2 = 58$

Rank Correlation between judges A and B is given by:

$$r_{AB} = 1 - \frac{6 \sum D_{AB}^2}{n(n^2-1)} = 1 - \frac{6(100)}{8(8^2-1)} = -0.1905$$

Rank Correlation between judges B and C is given by:

$$r_{BC} = 1 - \frac{6 \sum D_{BC}^2}{n(n^2-1)} = 1 - \frac{6(120)}{8(8^2-1)} = -0.4286$$

Rank Correlation between judges A and C is given by:

$$r_{AC} = 1 - \frac{6 \sum D_{AC}^2}{n(n^2-1)} = 1 - \frac{6(58)}{8(8^2-1)} = 0.3095$$

Therefore Judges A and C have the nearest approach to common tastes in beauty, while Judges B and C have most different beauty tastes.

Example 14: Obtain rank correlation coefficient for following marks in economics ($x$) and Mathematics ($y$) out of 25 for eight students.

<table>
<thead>
<tr>
<th>$x$</th>
<th>20</th>
<th>24</th>
<th>12</th>
<th>20</th>
<th>10</th>
<th>12</th>
<th>24</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y$</td>
<td>18</td>
<td>19</td>
<td>16</td>
<td>22</td>
<td>14</td>
<td>16</td>
<td>19</td>
<td>12</td>
</tr>
</tbody>
</table>

Solution: Converting data into ranks: Ranks of $x$ as $R_x$, Ranks of $y$ as $R_y$
Correction Factor = \( \sum m(m^2 - 1) \), \( m \) is the number of times each data value is repeated.
\[
\therefore \text{C. F.} = 2(2^2 - 1) + 3(3^2 - 1) + 2(2^2 - 1) + 2(2^2 - 1) + 2(2^2 - 1)
\]
\[
= 6 + 24 + 6 + 6 + 6 = 48
\]

Spearman’s Rank Correlation for repeated ranks is given by:
\[
r = 1 - \frac{6(\sum D^2 + \frac{1}{12}\text{C. F.})}{n(n^2 - 1)}, \text{ where } D = R_x - R_y
\]
\[
\therefore r = 1 - \frac{6(30 + \frac{48}{12})}{8(8^2 - 1)} = \frac{25}{42} = 0.595
\]

**Example 15** Obtain rank correlation coefficient for following data

<table>
<thead>
<tr>
<th></th>
<th>68</th>
<th>64</th>
<th>75</th>
<th>50</th>
<th>64</th>
<th>80</th>
<th>75</th>
<th>40</th>
<th>55</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62</td>
<td>58</td>
<td>68</td>
<td>45</td>
<td>81</td>
<td>60</td>
<td>68</td>
<td>48</td>
<td>50</td>
<td>70</td>
</tr>
</tbody>
</table>

**Solution:** Converting data into ranks: Ranks of \( x \) as \( R_x \), Ranks of \( y \) as \( R_y \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( R_x )</th>
<th>( Y )</th>
<th>( R_y )</th>
<th>( D = R_x - R_y )</th>
<th>( D^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>4</td>
<td>62</td>
<td>5</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>6</td>
<td>58</td>
<td>7</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>75</td>
<td>2.5</td>
<td>68</td>
<td>3.5</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>9</td>
<td>45</td>
<td>10</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>64</td>
<td>6</td>
<td>81</td>
<td>1</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>60</td>
<td>6</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>75</td>
<td>2.5</td>
<td>68</td>
<td>3.5</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>48</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>55</td>
<td>8</td>
<td>50</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>64</td>
<td>6</td>
<td>70</td>
<td>2</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

\( \sum D^2 = 72 \)
Correction Factor (C.F.) = $\sum m(m^2 - 1)$, \( m \) is the number of times each data value is repeated $\therefore$ C.F. = $2(2^2 - 1) + 3(3^2 - 1) + 2(2^2 - 1) = 36$

Spearman’s Rank Correlation for repeated ranks is given by:

$$ r = 1 - \frac{6(\sum D^2 + \frac{1}{12} \text{C.F.})}{n(n^2-1)}, \text{ where } D = R_x - R_y $$

$\therefore \ r = 1 - \frac{6(72 + \frac{36}{12})}{10(10^2-1)} = \frac{6}{11} = 0.545$

### 4.5 Linear Regression

Regression describes the functional relationship between dependent and independent variables; which helps us to make estimates of one variable from the other. Correlation quantifies the association between the two variables; whereas linear regression finds the best line that predicts \( y \) from \( x \) and also \( x \) from \( y \). The difference between correlation and regression is illustrated in the adjoining figure.

#### 4.5.1 Lines of Regression: If we plot the observations of the linear regression between two variables, actually two straight lines can approximately be drawn through the scatter diagram. One line estimates values of \( y \) for specified values of \( x \) (known as line of regression of \( y \) on \( x \)); and other predicts values of \( x \) from given values of \( y \) (called line of regression of \( x \) on \( y \)).

Let line of regression of \( y \) on \( x \) be represented by \( y = a + bx \) \( \quad \cdots \) ①

Normal equations as derived by the method of least square are:

$$ \sum y = an + b \sum x \quad \cdots \) ②$$

and

$$ \sum xy = a \sum x + b \sum x^2 \quad \cdots \) ③$$

Dividing ② by \( n \), we get

$$ \frac{\sum y}{n} = a + \frac{b \sum x}{n} \Rightarrow \bar{y} = a + b\bar{x} $$
Where $\bar{x}$ and $\bar{y}$ are the means of $x$ series and $y$ series. This shows that $(\bar{x}, \bar{y})$ lies on the line of regression given by (1).

Again as $(\bar{x}, \bar{y})$ satisfies (1), shifting the origin to $(\bar{x}, \bar{y})$ in equation (3), we get

$$\sum(x - \bar{x})(y - \bar{y}) = a \sum(x - \bar{x}) + b \sum(x - \bar{x})^2$$

$$\Rightarrow \sum(x - \bar{x})(y - \bar{y}) = b \sum(x - \bar{x})^2 \quad \therefore \sum(x - \bar{x}) = 0$$

$$\Rightarrow b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{\sum d_x d_y}{\sum d_x^2} \quad \ldots (4)$$

Again $r = \frac{\sum d_x d_y}{\sqrt{(\sum d_x^2)(\sum d_y^2)}} = \frac{\sum d_x d_y}{n \sqrt{\frac{\sum d_x^2}{n} \frac{\sum d_y^2}{n}}} = \frac{\sum d_x d_y}{n \sigma_x \sigma_y} \quad \therefore \sigma_x = \sqrt{\frac{\sum d_x^2}{n}}, \sigma_y = \sqrt{\frac{\sum d_y^2}{n}}$

Here $\sigma_x$, $\sigma_y$ are standard deviations of $x$ and $y$ data points respectively

$$\Rightarrow \sum d_x d_y = nr\sigma_x \sigma_y \quad \ldots (5)$$

Using (5) in (4), we get

$$b = \frac{nr \sigma_x \sigma_y}{\sum d_x^2} = \frac{r \sigma_x \sigma_y}{\sigma_x^2}$$

$$\Rightarrow b = \frac{r \sigma_y}{\sigma_x} \quad \text{which is slope of line of regression of $y$ on $x$}$$

$$\therefore b_{yx} = \frac{r \sigma_y}{\sigma_x}, \quad b_{yx} \text{ denotes slope of line of regression line of $y$ on $x$.}$$

Thus line of regression of $y$ on $x$ given by (1), passes through $(\bar{x}, \bar{y})$ and is having slope $b_{yx} = \frac{r \sigma_y}{\sigma_x}$

$$\therefore \text{Equation of line of regression of $y$ on $x$ is given by } y - \bar{y} = b_{yx}(x - \bar{x})$$

Similarly line of regression of $x$ on $y$ is given by: $x - \bar{x} = b_{xy}(y - \bar{y})$

where $b_{xy} = \frac{r \sigma_x}{\sigma_y}$ is slope of line of regression line of $x$ on $y$

Here $b_{xy}$ and $b_{yx}$ are known coefficients of regression and are connected by the relation:

$$b_{xy}b_{yx} = \left(\frac{r \sigma_x}{\sigma_y}\right)\left(\frac{r \sigma_y}{\sigma_x}\right) = r^2$$

### 4.5.2 Properties of Regression Coefficients

- As $\sqrt{b_{xy}b_{yx}} = r$, the coefficient of correlation is the geometric mean between the two regression coefficients.
- Since $\frac{b_{xy} + b_{yx}}{2} \geq \sqrt{b_{xy}b_{yx}} = r$, $\therefore$ arithmetic mean of the two regression coefficients is greater than or equal to the correlation coefficient ($r$).
If there is a perfect correlation between the two variables under consideration, then \( b_{xy} = b_{yx} = r \); and the two lines of regression coincide. Converse is also true, i.e. if two lines of regression coincide, then there is a perfect correlation; \( r = \pm 1 \).

Since \( b_{xy}b_{yx} = r^2 > 0 \), the signs of both regression coefficients \( b_{xy} \) and \( b_{yx} \) and coefficient of correlation (\( r \)) must be same; either all three negative or all positive.

\[ \therefore b_{xy}b_{yx} = r^2 \leq 1, \] if one of the regression coefficients is greater than unity, other must be less than unity.

Point of intersection of two lines of regression is \((\bar{x}, \bar{y})\), Where \( \bar{x} \) and \( \bar{y} \) are the means of \( x \) series and \( y \) series.

If both lines of regression cut each other at right angle, there is no correlation between the two variables; i.e. \( r = 0 \).

**Example 16** Prove that arithmetic mean of coefficients of regression is greater than the coefficient of correlation.

**Solution:** We know that \( b_{xy} = \frac{r \sigma_x}{\sigma_y} \) and \( b_{yx} = \frac{r \sigma_y}{\sigma_x} \)

To prove \( \frac{b_{xy} + b_{yx}}{2} > r \)

or \( \frac{1}{2} \left[ \frac{r \sigma_x}{\sigma_y} + \frac{r \sigma_y}{\sigma_x} \right] > r \)

or \( \frac{1}{2} \left[ \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y} \right] > 1 \)

or \( \left[ \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \sigma_y} \right] - 2 > 0 \)

or \( \sigma_x^2 + \sigma_y^2 - 2\sigma_x \sigma_y > 0 \)

or \( [\sigma_x - \sigma_y]^2 > 0 \)

which is true

Note: A.M. = \( r \) if \( b_{xy} = b_{yx} = r = \pm 1 \)

### 4.5.3 Angle between the Lines of Regression

If \( \theta \) be the acute angle between the two regression lines for two variables \( x \) and \( y \), then \( \tan \theta = \frac{1 - r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \)

**Proof:** The two lines of regression are given by:

\[ y - \bar{y} = \frac{r \sigma_y}{\sigma_x} (x - \bar{x}) \] ... \( 1 \)

and \( x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y}) \) ... \( 2 \)
If \( m_1 \) and \( m_2 \) are slopes of lines \( 1 \) and \( 2 \), then

\[
\tan \theta = \frac{m_2 - m_1}{1 + m_1 m_2}, \quad \text{where} \quad m_1 = \frac{r \sigma_y}{\sigma_x}, \quad m_2 = \frac{\sigma_y}{r \sigma_x}
\]

\[
\Rightarrow \tan \theta = \frac{\sigma_y - r \sigma_y}{1 + \frac{\sigma_y}{\sigma_x} \frac{\sigma_y}{\sigma_x}} = \frac{(1-r)^2 \sigma_y^2}{1 + \frac{\sigma_y^2}{\sigma_x^2} + \frac{\sigma_y^2}{\sigma_x^2}} = \frac{1-r^2}{r} \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \quad \ldots \,(3)
\]

\( \Rightarrow \) When \( r = 0 \), \( \tan \theta = \infty \Rightarrow \theta = \frac{\pi}{2} \) from \((3)\)

\( \therefore \) when \( r = 0 \), the two lines of regression are perpendicular to each other.

\( \Rightarrow \) When \( r = \pm 1 \), \( \tan \theta = 0 \Rightarrow \theta = 0 \) from \((3)\)

\( \therefore \) when \( r = \pm 1 \), the two lines of regression are coincident.

**Example 17** Find the correlation coefficient between \( x \) and \( y \), when the two lines of regression are given by: \( 2x - 9y + 6 = 0 \) and \( x - 2y + 1 = 0 \)

**Solution:** Let the line of regression of \( x \) on \( y \) be \( 2x - 9y + 6 = 0 \) \( \ldots (1) \)

Then the line of regression of \( y \) on \( x \) is \( x - 2y + 1 = 0 \) \( \ldots (2) \)

Now \((1) \Rightarrow x = \frac{9}{2}y - 3 \quad \therefore b_{xy} = \frac{9}{2} \)

Also \((2) \Rightarrow y = \frac{1}{2}x + \frac{1}{2} \quad \therefore b_{yx} = \frac{1}{2} \)

\( \therefore r = \sqrt{b_{xy} b_{yx}} = \sqrt{\frac{9}{2} \times \frac{1}{2}} = \frac{3}{2} \), which is not possible as \(-1 \leq r \leq 1\)

So our choice of regression lines is incorrect.

\( \therefore \) Line of regression of \( x \) on \( y \) is \( x = 2y - 1 \quad \therefore b_{xy} = 2 \)

Also line of regression of \( y \) on \( x \) is \( 2x - 9y + 6 = 0 \)

\( \Rightarrow y = \frac{2}{9}x + \frac{2}{3} \quad \therefore b_{yx} = \frac{2}{9} \)

\( \therefore r = \sqrt{b_{xy} b_{yx}} = \sqrt{2 \times \frac{2}{9}} = \frac{2}{3} \)

Hence coefficient of correlation between \( x \) and \( y \) is \( \frac{2}{3} \)

**Example 18** The regression equations calculated from a given set of observations for two random variables are: \( x = -0.4y + 6.4 \) and \( y = -0.6x + 4.6 \)

Calculate \( \bar{x}, \bar{y} \) and \( r \).

**Solution:** The two equations of regression are:

\( x = -0.4y + 6.4 \quad \ldots (1) \)
We know that the signs of \( b_{xy} \), \( b_{yx} \) and \( r \) must be same
\[ r = \pm 0.49 \]
Again we know that the point of intersection of two regression lines is \((\bar{x}, \bar{y})\)
Therefore solving (1) and (2), we get \(\bar{x} = 6, \bar{y} = 1\)

**Example 19** From a partially destroyed lab data, following results were retrieved:

Lines of regression are:
\[ x = 0.45y + 5.35 \quad \text{and} \quad y = 0.8x + 6.6, \quad \sigma_x^2 = 9 \]
Find \(\bar{x}, \bar{y}, \sigma_y\) and \(r\) for the existing data.

**Solution:** The two equations of regression are:
\[
\begin{align*}
\text{(1)} & \quad x = 0.45y + 5.35 \\
\text{(2)} & \quad y = 0.8x + 6.6
\end{align*}
\]
We know that the point of intersection of two regression lines is \((\bar{x}, \bar{y})\)
Therefore solving (1) and (2), we get \(\bar{x} = 13, \bar{y} = 17\)
Again \(\text{ (1)} \Rightarrow b_{xy} = 0.45 \quad \text{and} \quad b_{yx} = 0.8\)
\[ r^2 = b_{xy}b_{yx} = 0.36 \]
\[ r = \pm 0.6 \]
We know that the signs of \( b_{xy}, b_{yx} \) and \( r \) must be same
\[ r = 0.6 \]
Also \( b_{yx} = \frac{r \sigma_y}{\sigma_x} \Rightarrow 0.8 = \frac{(0.6)\sigma_y}{3} \Rightarrow \sigma_y = \frac{0.8 \times 3}{0.6} = 4 \)

**Example 20** Find the regression line of \(y\) on \(x\) from the following data:

<table>
<thead>
<tr>
<th>(x)</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>9</th>
<th>11</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>

Also estimate the value of \(y\), when \(x = 10\)

**Solution:** Let line of regression of \(y\) on \(x\) be:
\[ y = a + bx \quad \text{... (1)} \]
Then normal equations are given by:
\[
\sum y = an + b \sum x \quad \ldots \ (2)
\]

and \( \sum xy = a \sum x + b \sum x^2 \quad \ldots \ (3) \)

Calculating \( \sum x, \sum y, \sum xy \) and \( \sum x^2 \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x^2 )</th>
<th>( xy )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>81</td>
<td>63</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>121</td>
<td>88</td>
</tr>
<tr>
<td>14</td>
<td>9</td>
<td>196</td>
<td>126</td>
</tr>
</tbody>
</table>

\( \sum x = 56 \quad \sum y = 40 \quad \sum x^2 = 524 \quad \sum xy = 364 \)

Substituting values of \( \sum x, \sum y, \sum xy \) and \( \sum x^2 \) in (2) and (3)

\( \Rightarrow 40 = 8a + 56b \quad \ldots \ (4) \)

and \( 364 = 56a + 524b \quad \ldots \ (5) \)

Solving (4) and (5), we get \( a = \frac{6}{11} \) and \( b = \frac{7}{11} \)

Substituting in (1), line of regression of \( y \) on \( x \) is \( y = \frac{6}{11} + \frac{7}{11} x \)

\( \Rightarrow 7x - 11y + 6 = 0 \)

Also at \( x = 10 \), \( y = \frac{76}{11} \)

**Example 21** Following data depicts the statistical values of rainfall and production of wheat in a region for a specified time period.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production of Wheat (kg. per unit area)</td>
<td>10</td>
</tr>
<tr>
<td>Rainfall (cm)</td>
<td>8</td>
</tr>
</tbody>
</table>

Estimate the production of wheat when rainfall is 9cm if correlation coefficient between production and rainfall is given to be 0.5.

**Solution:** Let the variables \( x \) and \( y \) denote production and rainfall respectively.
Given that \( \bar{x} = 10 \), \( \bar{y} = 8 \) also \( \sigma_x = 8 \), \( \sigma_y = 2 \)

Now equation of regression of \( x \) on \( y \) is given by:

\[
x - \bar{x} = \frac{r \sigma_x}{\sigma_y} (y - \bar{y})
\]

\[
\Rightarrow x - 10 = \frac{0.5 \times 8}{2} (y - 8)
\]

\[
\Rightarrow x = 2y - 6
\]

\( \therefore \) When rainfall is 9cm, production of wheat is estimated to be

\( 2(9) - 6 = 12 \) kg. per unit area

**Example 22** Find the coefficient of correlation and the lines of regression for the data given below:

\( n = 18 \), \( \Sigma x = 12 \), \( \Sigma y = 18 \), \( \Sigma x^2 = 60 \), \( \Sigma y^2 = 96 \) and \( \Sigma xy = 48 \)

**Solution:**

\[
\bar{x} = \frac{\Sigma x}{n} = \frac{12}{18} = 0.67 \quad \bar{y} = \frac{\Sigma y}{n} = \frac{18}{18} = 1
\]

\[
\sigma_x^2 = \frac{\Sigma x^2}{n} - \left( \frac{\Sigma x}{n} \right)^2 = \frac{60}{18} - \left( \frac{12}{18} \right)^2 = 2.89 \quad \therefore \sigma_x = 1.7
\]

\[
\sigma_y^2 = \frac{\Sigma y^2}{n} - \left( \frac{\Sigma y}{n} \right)^2 = \frac{96}{18} - \left( \frac{18}{18} \right)^2 = 4.33 \quad \therefore \sigma_y = 2.08
\]

\[
r = \frac{\Sigma xy - \frac{1}{n} \Sigma x \Sigma y}{\sqrt{(\Sigma x^2 - \frac{1}{n} (\Sigma x)^2)} \sqrt{(\Sigma y^2 - \frac{1}{n} (\Sigma y)^2)}}
\]

\[
= \frac{48 - \frac{12 \times 18}{18}}{\sqrt{(60 - \frac{1}{18} 12^2)} \sqrt{(96 - \frac{1}{18} 18^2)}} = \frac{36}{(7.2)(8.83)} = 0.57
\]

\[
b_{yx} = \frac{r \sigma_x}{\sigma_y} = \frac{(0.57)(1.7)}{2.08} = 0.47 \quad b_{xy} = \frac{r \sigma_y}{\sigma_x} = \frac{(0.57)(2.08)}{1.7} = 0.7
\]

Equations of lines of regression are:

\( y - \bar{y} = b_{yx} (x - \bar{x}) \), \( x - \bar{x} = b_{xy} (y - \bar{y}) \)

\( \Rightarrow y - 1 = 0.7(x - 0.67) \) and \( x - 0.67 = 0.47(y - 1) \)

\( \Rightarrow y = 0.7x + 0.53 \) and \( x = 0.47y + 0.2 \)

**Example 23** Marks obtained by 11 students in statistics papers are given below:

<table>
<thead>
<tr>
<th>Paper I</th>
<th>60</th>
<th>65</th>
<th>68</th>
<th>70</th>
<th>75</th>
<th>85</th>
<th>80</th>
<th>45</th>
<th>55</th>
<th>56</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper II</td>
<td>62</td>
<td>64</td>
<td>65</td>
<td>70</td>
<td>74</td>
<td>90</td>
<td>82</td>
<td>56</td>
<td>50</td>
<td>48</td>
<td>60</td>
</tr>
</tbody>
</table>

Calculate the coefficient of correlation for the above data. Also find the equations of lines of regression.
**Solution:** Let marks obtained in paper I be denoted by \( x \) and marks obtained in paper II be denoted by \( y \).

Let \( A_x = 65, \ A_y = 70 \) \( \therefore \ d_x = x - 65, \ d_y = y - 70 \)

Calculating \( \sum d_x, \sum d_y, \sum d_x^2, \sum d_y^2 \) and \( \sum d_x d_y \)

<table>
<thead>
<tr>
<th>( x )</th>
<th>( d_x )</th>
<th>( d_x^2 )</th>
<th>( y )</th>
<th>( d_y )</th>
<th>( d_y^2 )</th>
<th>( d_x d_y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>-5</td>
<td>25</td>
<td>62</td>
<td>-8</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>65</td>
<td>0</td>
<td>0</td>
<td>64</td>
<td>-6</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>68</td>
<td>3</td>
<td>9</td>
<td>65</td>
<td>-5</td>
<td>25</td>
<td>-15</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>25</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>75</td>
<td>10</td>
<td>100</td>
<td>74</td>
<td>4</td>
<td>16</td>
<td>40</td>
</tr>
<tr>
<td>85</td>
<td>20</td>
<td>400</td>
<td>90</td>
<td>20</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>80</td>
<td>15</td>
<td>225</td>
<td>82</td>
<td>12</td>
<td>144</td>
<td>180</td>
</tr>
<tr>
<td>45</td>
<td>-20</td>
<td>400</td>
<td>56</td>
<td>-14</td>
<td>196</td>
<td>280</td>
</tr>
<tr>
<td>55</td>
<td>-10</td>
<td>100</td>
<td>50</td>
<td>-20</td>
<td>400</td>
<td>200</td>
</tr>
<tr>
<td>56</td>
<td>-9</td>
<td>81</td>
<td>48</td>
<td>-22</td>
<td>484</td>
<td>198</td>
</tr>
<tr>
<td>58</td>
<td>-7</td>
<td>49</td>
<td>60</td>
<td>-10</td>
<td>100</td>
<td>70</td>
</tr>
</tbody>
</table>

\[ \sum d_x = 2 \quad \sum d_x^2 = 1414 \quad \sum d_y = -49 \quad \sum d_y^2 = 1865 \quad \sum d_x d_y = 1393 \]

Karl Pearson coefficient of correlation \( (r) \) is given by:

\[
r = \frac{\sum d_x d_y - \frac{\Sigma d_x \Sigma d_y}{n}}{\sqrt{(\sum d_x^2 - \frac{1}{n} (\Sigma d_x)^2)(\sum d_y^2 - \frac{1}{n} (\Sigma d_y)^2)}}
\]

\[
\therefore r = \frac{1393 - \frac{(-2)(-49)}{11}}{\sqrt{(1414 - \frac{1}{11} (2)^2)(1865 - \frac{1}{11} (-49)^2)}} = \frac{1401.9091}{(37.5984)(40.5799)} = 0.9188
\]

Now \( \bar{x} = A_x + \frac{\sum d_x}{n} = 65 + \frac{2}{11} = 65.1818 \)

\( \bar{y} = A_y + \frac{\sum d_y}{n} = 70 + \frac{-49}{11} = 65.5455 \)

Also \( \sigma_x = \sqrt{\frac{\sum d_x^2}{n} - \left(\frac{\Sigma d_x}{n}\right)^2} = \sqrt{\frac{1414}{11} - \left(\frac{2}{11}\right)^2} = 11.3363 \)

and \( \sigma_y = \sqrt{\frac{\sum d_y^2}{n} - \left(\frac{\Sigma d_y}{n}\right)^2} = \sqrt{\frac{1865}{11} - \left(\frac{-49}{11}\right)^2} = 12.2353 \)
Equations of lines of regression are:

\[ y - \bar{y} = b_{yx} (x - \bar{x}) \quad , \quad x - \bar{x} = b_{xy} (y - \bar{y}) \]

\[ \Rightarrow y - 65.55 = 0.99(x - 65.18) \quad \text{and} \quad x - 65.18 = 0.85(y - 65.55) \]

\[ \Rightarrow y = 0.99x + 1.02 \quad \text{and} \quad x = 0.85y + 9.46 \]

**Example 24** The regression equations calculated from a given set of observations for two variables \( x \) and \( y \) are: \( x = 9y + 5 \) and \( y = kx + 9 \)

Show that \( 0 < k < \frac{1}{9} \). Also if \( k = \frac{1}{10} \), find \( \bar{x}, \bar{y} \) and \( r \)

**Solution:** The two equations of regression are:

\[ x = 9y + 5 \quad \ldots (1) \]

\[ y = kx + 9 \quad \ldots (2) \]

\[ \Rightarrow b_{xy} = 9 \quad \text{and} \quad b_{yx} = k \]

\[ \therefore r^2 = b_{xy} b_{yx} = 9k \]

\[ \Rightarrow r = 3\sqrt{k} \quad \therefore b_{xy} = 9 \quad \text{is positive, therefore} \quad k \quad \text{and} \quad r \quad \text{are also positive} \]

Now \( 0 < r < 1 \) or \( 0 < 3\sqrt{k} < 1 \)

\[ \Rightarrow 0 < 9k < 1 \quad \text{or} \quad 0 < k < \frac{1}{9} \]

Now if \( k = \frac{1}{10} \), equation (2) becomes \( 10y = x + 90 \quad \ldots (3) \)

Solving (1) and (3), the point of intersection of two regression lines is

\[ \bar{x} = 860 \quad , \quad \bar{y} = 95 \quad , \text{also} \quad r = 3\sqrt{k} = 3 \left( \frac{1}{\sqrt{10}} \right) = 0.949 \]

**Exercise 4**

1. Fit a straight line \( y = ax + b \) to the following data

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>3</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

2. Fit a straight line \( y = a + bx \) to the following data

<table>
<thead>
<tr>
<th></th>
<th>25</th>
<th>19</th>
<th>50</th>
<th>36</th>
<th>40</th>
<th>45</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>77</td>
<td>76</td>
<td>85</td>
<td>80</td>
<td>82</td>
<td>83</td>
<td>79</td>
</tr>
</tbody>
</table>
3. Fit a second degree parabola \( y = ax^2 + bx + c \) to the following data

\[
\begin{array}{c|cccc}
 x & 0 & 1 & 2 & 3 \\
 y & 1 & 1.8 & 1.3 & 2.5 \\
\end{array}
\]

4. Fit a second degree parabola \( y = a + bx + cx^2 \) to the following data

\[
\begin{array}{c|ccccc}
 x & 2 & 4 & 6 & 8 & 10 \\
 y & 3.07 & 12.85 & 31.47 & 57.38 & 91.29 \\
\end{array}
\]

5. Find the coefficient of correlation between \( x \) and \( y \) from the given data. Also find the two lines of regression.

\[
\begin{array}{c|ccccccccccc}
 x & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
 y & 10 & 12 & 16 & 28 & 25 & 36 & 41 & 49 & 40 & 50 \\
\end{array}
\]

6. Find the rank correlation for the following data:

\[
\begin{array}{c|cccccccccccccc}
 x & 56 & 42 & 72 & 36 & 63 & 47 & 55 & 49 & 38 & 42 & 68 & 60 \\
 y & 147 & 125 & 160 & 118 & 149 & 128 & 150 & 145 & 115 & 140 & 152 & 155 \\
\end{array}
\]

7. Following table shows ages of husband and wife of 53 married couples.

<table>
<thead>
<tr>
<th>Wife</th>
<th>husband↓</th>
<th>15-25</th>
<th>25-35</th>
<th>35-45</th>
<th>45-55</th>
<th>55-65</th>
<th>65-75</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-25</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25-35</td>
<td>2</td>
<td>12</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>35-45</td>
<td>-</td>
<td>4</td>
<td>10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>45-55</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>55-65</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>65-75</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
</tbody>
</table>

Calculate the coefficient of correlation between the age of the husband and that of wife.

8. The regression equations of two variables \( x \) and \( y \) are \( x = 0.7y + 5.2 \), \( y = 0.3x + 2.8 \). Find the means of the two variables and the coefficient of correlation between them.

9. If the coefficient of correlation between two variables \( x \) and \( y \) is 0.5 and the acute angle between their lines of regression is \( \tan^{-1} \frac{3}{8} \), show that \( \sigma_x = \frac{\sigma_y}{2 + \sqrt{3}} \).

10. From a partially destroyed lab data, following results were retrieved: Lines of regression are:

\[ 8x = 10y - 66 \text{ and } 18y = 40x - 214, \sigma_x = 3 \]

Find \( \bar{x}, \bar{y}, \sigma_y \) and \( r \) for the existing data.
Answers

1. \( y = 0.38x + 1.6 \)
2. \( y = 70.052 + 0.292x \)
3. \( y = 0.55x^2 - 1.07x + 1.42 \)
4. \( y = 0.34 - 0.78x + 0.99x^2 \)
5. \( r = 0.96, x = 0.2y - 0.64, y = 4.69x + 4.9 \)
6. 0.932
7. 0.91
8. \( \bar{x} = 9.06, \bar{y} = 5.52, r = 0.46 \)
10. \( \bar{x} = 13, \bar{y} = 17, \sigma_y = 4, r = 0.6 \)