CHAPTER 7: CORRELATION AND REGRESSION

Prof. Alan Wan
1. Introduction
   1.1 Linear Association

2. Covariance and Correlation
   2.1 Covariance
   2.2 Correlation

3. Regression
   3.1 Simple Linear Regression
   3.2 Multiple Linear Regression
In this chapter, we are concerned with two related problems: i) first, that of establishing whether or not two variables are related, and then, if they are, ii) using that relationship to forecast one variable for given values of the other.
In this chapter, we are concerned with two related problems: i) first, that of establishing whether or not two variables are related, and then, if they are, ii) using that relationship to forecast one variable for given values of the other.

Such forecast or prediction is useful because there are many situations where values of one variable are known or can easily be obtained while the other variable is unknown.
In this chapter, we are concerned with two related problems: i) first, that of establishing whether or not two variables are related, and then, if they are, ii) using that relationship to forecast one variable for given values of the other.

Such forecast or prediction is useful because there are many situations where values of one variable are known or can easily be obtained while the other variable is unknown.

For example, when the price of a product is changed, the new price is known but sales for the future periods are not.
Correlation coefficients are used to solve problem i), while regression can be used to cope with problem ii).
Correlation coefficients are used to solve problem i), while regression can be used to cope with problem ii).

The typical situation that we will look at relates to two variables measuring different characteristics of some population of individuals (although we will also touch on the case of three or more variables near the end of this chapter).

For the two-variable case, the available information will usually consist of paired sample data corresponding to pairs of observations on the two variables for $n$ members of a sample taken from the population.
If two variables are related, then the nature of the relationship may be indicated by plotting paired samples of observations for both variables on a scatter diagram.
Linear association

- If two variables are related, then the nature of the relationship may be indicated by plotting paired samples of observations for both variables on a scatter diagram.
- For example, consider the following data for variables $X=$ height (in cm.) and $Y=$ IQ from a sample of 10 students:

<table>
<thead>
<tr>
<th>$X$</th>
<th>170</th>
<th>185</th>
<th>165</th>
<th>140</th>
<th>180</th>
<th>150</th>
<th>200</th>
<th>160</th>
<th>175</th>
<th>170</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>120</td>
<td>130</td>
<td>140</td>
<td>135</td>
<td>100</td>
<td>115</td>
<td>130</td>
<td>125</td>
<td>145</td>
<td>110</td>
</tr>
</tbody>
</table>
Linear association

- The scatter diagram for these data is as follows (where each dot represents a student):

![Scatter diagram](image)

- This diagram indicates no obvious relationship between $X$ and $Y$, as you might well expect, since there is no known relationship between height and IQ.
However, consider the scatter diagram between height $X$ and weight $Z$ (in kg.) for the same ten students:

$X$  170  185  165  140  180  150  200  160  175  170
$Z$  75   80   75   50   70   60   100  65   80   70
However, consider the scatter diagram between height $X$ and weight $Z$ (in kg.) for the same ten students:

$X$  170 185 165 140 180 150 200 160 175 170  
$Z$  75  80  75  50  70  60 100  65  80  70

There is a clear tendency for small values of $X$ to be associated with small values of $Z$, and large $X$ with large $Z$. 
The scatter diagram is as follows:

The dots on the scatter diagram lie "close to" a straight line with a positive slope. We say that these two variables, height and weight, have a positive linear association.
For a different kind of relationship, consider the data from 10 firms on the variable $S =$ amount spent by the firm per employee on job safety education ($000) and $T =$ number of days lost through industrial accidents over one year.

$S \quad 5 \quad 0 \quad 5 \quad 1 \quad 4 \quad 8 \quad 7 \quad 2 \quad 6 \quad 3$

$T \quad 3 \quad 6 \quad 5 \quad 8 \quad 2 \quad 1 \quad 2 \quad 4 \quad 3 \quad 8$
Linear association

- For a different kind of relationship, consider the data from 10 firms on the variable $S =$ amount spent by the firm per employee on job safety education ($000) and $T =$ number of days lost through industrial accidents over one year.

\[
\begin{align*}
S & = 5\ 0\ 5\ 1\ 4\ 8\ 7\ 2\ 6\ 3 \\
T & = 3\ 6\ 5\ 8\ 2\ 1\ 2\ 4\ 3\ 8
\end{align*}
\]

- The prevailing tendency appears for small values of $S$ to be associated with large values of $T$ and vice versa. The dots lie "close to" a straight line with a negative slope. We say that the variables exhibit a **negative linear association**.
The scatter diagram looks like this:
1. Introduction
2. Covariance and Correlation
3. Regression

1.1 Linear Association

Linear association

- The scatter diagram looks like this:

![Scatter diagram](image)

- If two variables are linear associated, then it is quite easy, given paired sample data, to fit a straight line to the scatter diagram and use this to forecast values of one variable given values of the other. This is what we do in **linear regression**.
Question: So how do we measure the degree of linear association between two variables $X$ and $Y$?
Question: So how do we measure the degree of linear association between two variables $X$ and $Y$?

Answer: The covariance is the quantity that measures this linear association.
Question: So how do we measure the degree of linear association between two variables $X$ and $Y$?

Answer: The covariance is the quantity that measures this linear association.

Definition: The population covariance, denoted by $\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)$, measures the average value of $(X - \mu_X)(Y - \mu_Y)$ in the population.
Question: So how do we measure the degree of linear association between two variables $X$ and $Y$?

Answer: The covariance is the quantity that measures this linear association.

Definition: The population covariance, denoted by 
$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y),$$
measures the average value of $(X - \mu_X)(Y - \mu_Y)$ in the population.

Definition: The sample covariance, denoted by 
$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}),$$
is an estimator of $\sigma_{XY}$ based on $n$ pairs of sample values.
Clearly, the amounts \((X_i - \mu_X)\) and \((Y_i - \mu_Y)\) measure the amounts by which \(X\) and \(Y\) are above or below the average. If \(X_i\) and \(Y_i\) are both above or both below their respective averages, then the cross product term \((X_i - \mu_X)(Y_i - \mu_Y)\) will be positive; if this is generally true for the \(X\) and \(Y\) values in the population then the covariance \(\sigma_{XY}\) is positive. In such case, we say that \(X\) and \(Y\) are positively correlated.
Covariance

- Clearly, the amounts \((X_i - \mu_X)\) and \((Y_i - \mu_Y)\) measure the amounts by which \(X\) and \(Y\) are above or below the average. If \(X_i\) and \(Y_i\) are both above or both below their respective averages, then the cross product term \((X_i - \mu_X)(Y_i - \mu_Y)\) will be positive; if this is generally true for the \(X\) and \(Y\) values in the population then the covariance \(\sigma_{XY}\) is positive. In such case, we say that \(X\) and \(Y\) are positively correlated.

- Conversely, if \(X\) is above the average when \(Y\) is below the average or vice versa then the cross product term \((X_i - \mu_X)(Y_i - \mu_Y)\) will be negative; if this is generally true for the \(X\) and \(Y\) values in the population then the covariance \(\sigma_{XY}\) is negative. In such case, we say that \(X\) and \(Y\) are negatively correlated.
Covariance

- Clearly, the amounts \((X_i - \mu_X)\) and \((Y_i - \mu_Y)\) measure the amounts by which \(X\) and \(Y\) are above or below the average. If \(X_i\) and \(Y_i\) are both above or both below their respective averages, then the cross product term \((X_i - \mu_X)(Y_i - \mu_Y)\) will be positive; if this is generally true for the \(X\) and \(Y\) values in the population then the covariance \(\sigma_{XY}\) is positive. In such case, we say that \(X\) and \(Y\) are positively correlated.

- Conversely, if \(X\) is above the average when \(Y\) is below the average or vice versa then the cross product term \((X_i - \mu_X)(Y_i - \mu_Y)\) will be negative; if this is generally true for the \(X\) and \(Y\) values in the population then the covariance \(\sigma_{XY}\) is negative. In such case, we say that \(X\) and \(Y\) are negatively correlated.

- The same interpretations apply also to the sample covariance \(s_{XY}\).
Covariance

The cross product terms will be positive in quadrants I and III, and negative in quadrants II and IV. With positive linear association there is a tendency for the dots to lie predominantly in quadrants I and III, as illustrated below:
On the other hand, with negative linear association there is a tendency for the dots to lie predominantly in quadrants II and IV:
If there is no or very weak linear association then there is a tendency for the dots to scatter across all four quadrants:
The covariance only measures linear association - a covariance of zero does not necessarily imply that $X$ and $Y$ have no association because they may be related in a non-linear way; for example,
Example 1: Calculate $s_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>
Covariance

Example 1: Calculate $s_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

- $X - \bar{X}$
- $Y - \bar{Y}$
- $(X - \bar{X})(Y - \bar{Y})$

<table>
<thead>
<tr>
<th></th>
<th>$X - \bar{X}$</th>
<th>$Y$</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-3</td>
<td>1</td>
<td>-8</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>4</td>
<td>-5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>8</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>20</td>
<td>11</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>65</td>
</tr>
</tbody>
</table>

$\bar{X} = 15/5 = 3$, $\bar{Y} = 45/5 = 9$, $s_{XY} = 65/4 = 16.25$
Example 2: Calculate $s_{XY}$ for this sample data:

\[
\begin{array}{cccccc}
X & 0 & 2 & 3 & 4 & 6 \\
Y & 8 & 5 & 2 & -2 & -3 \\
\end{array}
\]
### Example 2: Calculate $s_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$X - \bar{X}$</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td>-3</td>
<td>6</td>
<td>-18</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>-2</td>
<td>1</td>
<td>-4</td>
<td>-4</td>
</tr>
<tr>
<td>6</td>
<td>-3</td>
<td>3</td>
<td>-5</td>
<td>-15</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>-40</td>
</tr>
</tbody>
</table>

- $\bar{X} = 15/5 = 3$, $\bar{Y} = 10/5 = 2$, $s_{XY} = -40/4 = -10$
Example 3: Suppose that values of $X$ and $Y$ in Example 2 are measured in dollars. Now, let’s convert the dollars into cents:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>80</td>
<td>50</td>
<td>20</td>
<td>-20</td>
<td>-30</td>
</tr>
</tbody>
</table>
Example 3: Suppose that values of $X$ and $Y$ in Example 2 are measured in dollars. Now, let’s convert the dollars into cents:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0  20  30  40  60</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>80  50  20 -20 -30</td>
</tr>
</tbody>
</table>

\[
egin{array}{cccccc}
X & X - \bar{X} & Y & Y - \bar{Y} & (X - \bar{X})(Y - \bar{Y}) \\
0 & -30 & 80 & 60 & -1800 \\
20 & -10 & 50 & 30 & -300 \\
30 & 0  & 20 & 0  & 0 \\
40 & 10 & -20 & -40 & -400 \\
60 & 30 & -30 & -50 & -1500 \\
15 & 0  & 100 & 0  & -4000 \\
\end{array}
\]

$\bar{X} = 150/5 = 30$, $\bar{Y}=100/5=20$, $s_{XY}=-4000/4=-1000$
Example 3: Suppose that values of $X$ and $Y$ in Example 2 are measured in dollars. Now, let’s convert the dollars into cents:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>80</td>
<td>50</td>
<td>20</td>
<td>-20</td>
<td>-30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$X - \bar{X}$</th>
<th>$Y$</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-30</td>
<td>80</td>
<td>60</td>
<td>-1800</td>
</tr>
<tr>
<td>20</td>
<td>-10</td>
<td>50</td>
<td>30</td>
<td>-300</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>-20</td>
<td>-40</td>
<td>-400</td>
</tr>
<tr>
<td>60</td>
<td>30</td>
<td>-30</td>
<td>-50</td>
<td>-1500</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>-4000</td>
</tr>
</tbody>
</table>

$\bar{X} = 150/5 = 30, \quad \bar{Y}=100/5=20, \quad s_{XY}=-4000/4=-1000$

The sample covariance is inflated by a factor of 100 although the relationship between $X$ and $Y$ is really the same as in the last example.
Correlation

▶ One problem with the covariance is that it is dependent on the units used to measure $X$ and $Y$ and thus usually cannot be directly compared for different variables. This scaling problem affects the magnitude (but not the sign) of the linear association between $X$ and $Y$ as reported by the covariance.
Correlation

- One problem with the covariance is that it is dependent on the units used to measure $X$ and $Y$ and thus usually cannot be directly compared for different variables. This scaling problem affects the magnitude (but not the sign) of the linear association between $X$ and $Y$ as reported by the covariance.

- The **correlation coefficient** is a measure of the linear association between two variables that is *not* affected by the variables’ units of measure. It *adjusts* the covariance by the standard deviations of $X$ and $Y$ so that the resulting measure is unit-free.
Correlation

- One problem with the covariance is that it is dependent on the units used to measure X and Y and thus usually cannot be directly compared for different variables. This scaling problem affects the magnitude (but not the sign) of the linear association between X and Y as reported by the covariance.

- The correlation coefficient is a measure of the linear association between two variables that is not affected by the variables’ units of measure. It adjusts the covariance by the standard deviations of X and Y so that the resulting measure is unit-free.

- The correlation coefficient is a ”standardised score” of the covariance.
Definition: The population correlation coefficient is defined as
\[
\rho_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2 \sum_{i=1}^{N}(Y_i - \mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}
\]
Definition: The **population correlation coefficient** is defined as

$$\rho_{XY} = \frac{\sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N} (X_i - \mu_X)^2} \sqrt{\sum_{i=1}^{N} (Y_i - \mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Definition: The **sample correlation coefficient** is defined as

$$r_{XY} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{s_{XY}}{s_X s_Y}$$
Correlation

Definition: The population correlation coefficient is defined as
\[ \rho_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2 \sum_{i=1}^{N}(Y_i - \mu_Y)^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

Definition: The sample correlation coefficient is defined as
\[ r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{s_{XY}}{s_X s_Y} \]

Notice that the denominator of \( \rho_{XY} \) is always non-negative as it is a product of two standard deviation measures. Hence the sign of \( \rho_{XY} \) is the same as that of \( \sigma_{XY} \). Similarly, the sign of \( r_{XY} \) is the same as that of \( s_{XY} \).
Correlation

- It can be shown it is always the case that $-1 \leq \rho_{XY} \leq 1$ and $-1 \leq r_{XY} \leq 1$
Correlation

- It can be shown it is always the case that $-1 \leq \rho_{XY} \leq 1$ and $-1 \leq r_{XY} \leq 1$

- Three special values of $\rho_{XY}$ and $r_{XY}$ are of interest:
  1. When $\rho_{XY} = 0$ ($r_{XY} = 0$), $X$ and $Y$ are not linearly related, and we say that $X$ and $Y$ are uncorrelated in the population (sample);
  2. When all population (sample) values of $X$ and $Y$ lie exactly on a straight line having a positive slope (e.g., $Y = a + bX$, with $b > 0$), then $\rho_{XY} = 1$ ($r_{XY} = 1$);
  3. When all population (sample) values of $X$ and $Y$ lie exactly on a straight line having a negative slope (e.g., $Y = a + bX$, with $b < 0$), then $\rho_{XY} = -1$ ($r_{XY} = -1$).
Correlation

- It can be shown it is always the case that 
  \[-1 \leq \rho_{XY} \leq 1 \text{ and } -1 \leq r_{XY} \leq 1\]

- Three special values of \( \rho_{XY} \) and \( r_{XY} \) are of interest:
  1. When \( \rho_{XY} = 0 \) (\( r_{XY} = 0 \)), \( X \) and \( Y \) are not linearly related, and we say that \( X \) and \( Y \) are uncorrelated in the population (sample);
  2. When all population (sample) values of \( X \) and \( Y \) lie exactly on a straight line having a positive slope (e.g., \( Y = a + bX \), with \( b > 0 \)), then \( \rho_{XY} = 1 \) (\( r_{XY} = 1 \));
  3. When all population (sample) values of \( X \) and \( Y \) lie exactly on a straight line having a negative slope (e.g., \( Y = a + bX \), with \( b < 0 \)), then \( \rho_{XY} = -1 \) (\( r_{XY} = -1 \))

- If the population (sample) values of \( X \) and \( Y \) lie close to a straight line, then \( \rho_{XY} \) (\( r_{XY} \)) will be close to 1 or -1.
The following interpretations may be applied to $\rho_{XY}$:

1. $|\rho_{XY}| \geq 0.9$, i.e., $\rho_{XY} \leq -0.9$ or $\rho_{XY} \geq 0.9$: this indicates a very strong linear association between the two variables in the population.

2. $|\rho_{XY}| \leq 0.2$, i.e., $-0.2 \leq \rho_{XY} \leq 0.2$: this indicates a weak linear association between the two variables in the population.

3. $0.2 \leq |\rho_{XY}| \leq 0.9$, i.e., $-0.9 \leq \rho_{XY} \leq -0.2$ or $0.2 \leq \rho_{XY} \leq 0.9$: there is a linear association between $X$ and $Y$ with moderate to high strength in the population.

The same interpretations can be applied to $r_{XY}$ by replacing "$\rho_{XY}$" by "$r_{XY}$" and "population" by "sample" in the above statements 1, 2 and 3.
Correlation

- The following interpretations may be applied to $\rho_{XY}$:
  1. $|\rho_{XY}| \geq 0.9$, i.e., $\rho_{XY} \leq -0.9$ or $\rho_{XY} \geq 0.9$:
     this indicates a very strong linear association between the two variables in the population.
Correlation

The following interpretations may be applied to $\rho_{XY}$:

1. $|\rho_{XY}| \geq 0.9$, i.e., $\rho_{XY} \leq -0.9$ or $\rho_{XY} \geq 0.9$:
   this indicates a very strong linear association between the two variables in the population.

2. $|\rho_{XY}| \leq 0.2$, i.e., $-0.2 \leq \rho_{XY} \leq 0.2$:
   this indicates a weak linear association between the two variables in the population.
Correlation

The following interpretations may be applied to $\rho_{XY}$:

1. $|\rho_{XY}| \geq 0.9$, i.e., $\rho_{XY} \leq -0.9$ or $\rho_{XY} \geq 0.9$:
   this indicates a very strong linear association between the two variables in the population.

2. $|\rho_{XY}| \leq 0.2$, i.e., $-0.2 \leq \rho_{XY} \leq 0.2$:
   this indicates a weak linear association between the two variables in the population.

3. $0.2 \leq |\rho_{XY}| \leq 0.9$, i.e., $-0.9 \leq \rho_{XY} \leq -0.2$ or $0.2 \leq \rho_{XY} \leq 0.9$:
   there is a linear association between $X$ and $Y$ with moderate to high strength in the population.
Correlation

The following interpretations may be applied to $\rho_{XY}$:

1. $|\rho_{XY}| \geq 0.9$, i.e., $\rho_{XY} \leq -0.9$ or $\rho_{XY} \geq 0.9$:
   this indicates a very strong linear association between the two variables in the population.

2. $|\rho_{XY}| \leq 0.2$, i.e., $-0.2 \leq \rho_{XY} \leq 0.2$:
   this indicates a weak linear association between the two variables in the population.

3. $0.2 \leq |\rho_{XY}| \leq 0.9$, i.e., $-0.9 \leq \rho_{XY} \leq -0.2$ or $0.2 \leq \rho_{XY} \leq 0.9$:
   there is a linear association between $X$ and $Y$ with moderate to high strength in the population.

The same interpretations can be applied to $r_{XY}$ by replacing "$\rho_{XY}$" by "$r_{XY}$" and "population" by "sample" in the above statements 1, 2 and 3.
Correlation

- Here are some diagrams illustrating different values of $r_{XY}$:
Example 4: Consider the data observations in Example 1 again. Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>20</td>
</tr>
</tbody>
</table>

$\bar{X} = \frac{15}{5} = 3, \quad \bar{Y} = \frac{45}{5} = 9,$

$s_{XY} = \frac{65}{4} = 16.25,$

$r_{XY} = \frac{65}{\sqrt{20} \times 220} = 0.9799$
Correlation

Example 4: Consider the data observations in Example 1 again. Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>$X - \bar{X}$</th>
<th>Y</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
<th>$(X - \bar{X})^2$</th>
<th>$(Y - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>1</td>
<td>-8</td>
<td>24</td>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>-1</td>
<td>4</td>
<td>-5</td>
<td>5</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>3</td>
<td>20</td>
<td>11</td>
<td>33</td>
<td>9</td>
<td>121</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>65</td>
<td>20</td>
<td>220</td>
</tr>
</tbody>
</table>

$\bar{X} = 15/5 = 3$, $\bar{Y} = 45/5 = 9$, $s_{XY} = 65/4 = 16.25$, $r_{XY} = 65 / \sqrt{20 \times 220} = 0.9799$
Example 5: Consider the data observations in Example 2 again. Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>8</td>
<td>5</td>
<td>2</td>
<td>-2</td>
<td>-3</td>
</tr>
</tbody>
</table>

$\bar{X} = \frac{15}{5} = 3$, $\bar{Y} = \frac{10}{5} = 2$, $s_{XY} = -\frac{40}{4} = -10$, $r_{XY} = -\frac{-40}{\sqrt{20} \times 86} = -0.9645$
Example 5: Consider the data observations in Example 2 again. Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>-2</td>
</tr>
<tr>
<td>6</td>
<td>-3</td>
</tr>
</tbody>
</table>

$\bar{X} = 15/5 = 3, \; \bar{Y} = 10/5 = 2, \; s_{XY} = -40/4 = -10,$

$$r_{XY} = -40/\sqrt{20 \times 86} = -0.9645$$
Example 6: Consider the data observations in Example 3 which are multiples of the data in Example 2 (and Example 5) by a factor of 10:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>80</td>
<td>50</td>
<td>20</td>
<td>-20</td>
<td>-30</td>
</tr>
</tbody>
</table>

Given 

\[
\bar{X} = \frac{150}{5} = 30, \quad \bar{Y} = \frac{100}{5} = 20,
\]

\[
s_{XY} = \frac{-4000}{4} = -1000,
\]

\[
r_{XY} = \frac{-4000}{\sqrt{2000} \times 8600} = -0.9645
\]

(i.e., \(r_{XY}\) remains unchanged although \(s_{XY}\) has been inflated by a factor of 100).
2. Covariance and Correlation

2.1 Covariance

2.2 Correlation

Example 6: Consider the data observations in Example 3 which are multiples of the data in Example 2 (and Example 5) by a factor of 10:

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>40</td>
<td>-20</td>
</tr>
<tr>
<td>60</td>
<td>-30</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccccccc}
X & X - \bar{X} & Y & Y - \bar{Y} & (X - \bar{X})(Y - \bar{Y}) & (X - \bar{X})^2 & (Y - \bar{Y})^2 \\
0 & -30 & 80 & 60 & -1800 & 900 & 3600 \\
20 & -10 & 50 & 30 & -300 & 100 & 900 \\
30 & 0 & 20 & 0 & 0 & 0 & 0 \\
40 & 10 & -20 & -40 & -400 & 100 & 1600 \\
60 & 30 & -30 & -50 & -1500 & 900 & 2500 \\
15 & 0 & 100 & 0 & -4000 & 2000 & 8600 \\
\end{array}
\]

\[\bar{X} = 150/5 = 30, \quad \bar{Y} = 100/5 = 20, \quad s_{XY} = -4000/4 = -1000, \]
\[r_{XY} = -4000/\sqrt{2000 \times 8600} = -0.9645 \text{ (i.e., } r_{XY} \text{ remains unchanged although } s_{XY} \text{ has been inflated by a factor of 100).}\]
Example 7: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>1000</td>
<td>-90</td>
<td>230</td>
<td>-200</td>
<td>900</td>
</tr>
</tbody>
</table>

$\bar{X} = \frac{150}{5} = 30$, $\bar{Y} = \frac{1840}{4} = 368$, $s_{XY} = -\frac{4100}{4} = -1025$, $r_{XY} = -\frac{4100}{\sqrt{2000} \times 1233880} = -0.082534$. 

Correlation
Example 7: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>$X - \bar{X}$</th>
<th>$Y$</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
<th>$(X - \bar{X})^2$</th>
<th>$(Y - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-30</td>
<td>1000</td>
<td>632</td>
<td>-18960</td>
<td>900</td>
<td>399424</td>
</tr>
<tr>
<td>20</td>
<td>-10</td>
<td>-90</td>
<td>-458</td>
<td>4580</td>
<td>100</td>
<td>209764</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>230</td>
<td>-138</td>
<td>0</td>
<td>0</td>
<td>19044</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>-200</td>
<td>-568</td>
<td>-5680</td>
<td>100</td>
<td>322624</td>
</tr>
<tr>
<td>60</td>
<td>30</td>
<td>900</td>
<td>-532</td>
<td>15960</td>
<td>900</td>
<td>283024</td>
</tr>
<tr>
<td>150</td>
<td>0</td>
<td>1840</td>
<td>0</td>
<td>-4100</td>
<td>2000</td>
<td>1233880</td>
</tr>
</tbody>
</table>

$\bar{X} = 150/5 = 30$, $\bar{Y} = 1840/4 = 368$, $s_{XY} = -4100/4 = -1025$, $r_{XY} = -4100/\sqrt{2000 \times 1233880} = -0.082534$. 
Example 8: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>5</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

$\bar{X} = \frac{15}{5} = 3$, $\bar{Y} = \frac{55}{5} = 11$, $s_{XY} = \frac{40}{4} = 10$,

$r_{XY} = \frac{40}{\sqrt{20} \times \sqrt{80}} = 1$ (i.e., perfect linear association)

Note that $Y = 5 + 2X$. 

$\frac{\text{Example 8 answer}}{\text{Example 8 answer}}$
Example 8: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>5</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$X - \bar{X}$</th>
<th>$Y$</th>
<th>$Y - \bar{Y}$</th>
<th>$(X - \bar{X})(Y - \bar{Y})$</th>
<th>$(X - \bar{X})^2$</th>
<th>$(Y - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-3</td>
<td>5</td>
<td>-6</td>
<td>18</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>9</td>
<td>-2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>17</td>
<td>6</td>
<td>18</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>55</td>
<td>0</td>
<td>40</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>

$\bar{X} = 15/5 = 3$, $\bar{Y} = 55/5 = 11$, $s_{XY} = 40/4 = 10$, $r_{XY} = 40 / \sqrt{20 \times 80} = 1$ (i.e., perfect linear association)

Note that $Y = 5 + 2X$
Correlation

Example 9: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-6</td>
</tr>
</tbody>
</table>

Calculate $r_{XY}$:

$\bar{X} = \frac{15}{5} = 3$

$\bar{Y} = \frac{-15}{5} = -3$

$s_{XY} = \frac{-20}{4} = -5$

$r_{XY} = \frac{-20}{\sqrt{20} \times 20} = -1$ (i.e., perfect linear association)
Example 9: Calculate $r_{XY}$ for this sample data:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>$X-ar{X}$</th>
<th>$Y$</th>
<th>$Y-ar{Y}$</th>
<th>$(X-ar{X})(Y-ar{Y})$</th>
<th>$(X-ar{X})^2$</th>
<th>$(Y-ar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>3</td>
<td>-9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>-4</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>-6</td>
<td>-3</td>
<td>-9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>-15</td>
<td>0</td>
<td>-20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

$\bar{X} = 15/5 = 3$, $\bar{Y} = -15/5 = -3$, $s_{XY} = -20/4 = -5$,

$r_{XY} = -20/\sqrt{20 \times 20} = -1$ (i.e., perfect linear association)

Note that $Y = -X$
Correlation

- The sample correlation coefficient $r_{XY}$ is just an estimator of $\rho_{XY}$; $r_{XY}$ is meaningful only for what it tells you about the population correlation coefficient, because it is $\rho_{XY}$ that measures the true relationship between $X$ and $Y$, where $r_{XY}$ relates only to the particular sample values observed.
Correlation

- The sample correlation coefficient $r_{XY}$ is just an estimator of $\rho_{XY}$; $r_{XY}$ is meaningful only for what it tells you about the population correlation coefficient, because it is $\rho_{XY}$ that measures the true relationship between $X$ and $Y$, where $r_{XY}$ relates only to the particular sample values observed.

- More precisely, the value of $r_{XY}$ may be used to test certain hypothesis about the population correlation coefficient $\rho_{XY}$. If $X$ and $Y$ have a joint normal distribution then the following statistic $t$ may be used to test the hypothesis:

$$H_0 : \rho_{XY} = 0 \text{ vs. } H_1 : \rho_{XY} \neq 0$$

$$t = \frac{r_{XY} \sqrt{n-2}}{\sqrt{(1-r_{XY}^2)}} \sim t_{n-2}$$
Example 10: Suppose based on $n = 20$ pairs of observations on $X$ and $Y$ we obtain $r_{XY} = 0.6$ and want to test

$H_0 : \rho_{XY} = 0$ vs. $H_1 : \rho_{XY} \neq 0$

at $\alpha = 0.05$. 

$\text{Now, } t = 0.6 \sqrt{20 - 2} \sqrt{1 - r_{XY}^2} = 3.18$, and $t(0.05/2, 18) = 2.101$.

Hence we reject $H_0$ and conclude that there is a significant linear association between $X$ and $Y$. 


Example 10: Suppose based on $n = 20$ pairs of observations on $X$ and $Y$ we obtain $r_{XY} = 0.6$ and want to test

$$H_0 : \rho_{XY} = 0 \text{ vs. } H_1 : \rho_{XY} \neq 0$$

at $\alpha = 0.05$.

Now, $t = \frac{0.6\sqrt{20-2}}{\sqrt{(1-0.6^2)}} = 3.18$, and $t(0.05/2,18) = 2.101$

Hence we reject $H_0$ and conclude that there is a significant linear association between $X$ and $Y$. 
Suppose that a scatter diagram indicates a linear association between two variables, i.e., the points corresponding to paired observations lie close to a straight line. Then it would make sense to use the line to predict one variable from given values of the other. But how do we know which line to choose?
Simple Linear Regression

- Suppose that a scatter diagram indicates a linear association between two variables, i.e., the points corresponding to paired observations lie close to a straight line. Then it would make sense to use the line to predict one variable from given values of the other. But how do we know which line to choose?

- It is possible to fit a line by "eye", but many possible lines will look equally good.
Simple Linear Regression

To single out one line we need to specify a criterion for best fit. Suppose that $Y$, the variable that we want to predict, is plotted on the vertical axis. The rationale that has been adopted is to choose the line that minimises the sum of squared differences between the observed $Y$ values and the values on the fitted line, i.e., the sum of the squared vertical deviations between that line and the sample points (the dotted lengths in the diagram).
Simple Linear Regression
Suppose that the true relationship between $Y$ and $X$ is

$$Y = \beta_0 + \beta_1 X + u,$$

where $u$ is an unobserved disturbance term and $\beta_0$ and $\beta_1$ are unknown coefficients to be estimated.
Simple Linear Regression

- Suppose that the true relationship between $Y$ and $X$ is
  $$Y = \beta_0 + \beta_1 X + u,$$
  where $u$ is an unobserved disturbance term and $\beta_0$ and $\beta_1$ are unknown coefficients to be estimated.

- $Y$ is called a response or dependent variable and $X$ is an explanatory variable as it explains the behaviour of $Y$. 
Simple Linear Regression

- Suppose that the true relationship between $Y$ and $X$ is
  \[ Y = \beta_0 + \beta_1 X + u, \]
  where $u$ is an unobserved disturbance term and $\beta_0$ and $\beta_1$ are unknown coefficients to be estimated.

- $Y$ is called a **response or dependent variable** and $X$ is an **explanatory variable** as it **explains** the behaviour of $Y$.

- This model is known as a **simple linear regression**, as it postulates, first, a linear statistical relationship between $Y$ and $X$, and that second, $X$ is the only explanatory variable in the study; the model becomes a multiple linear regression (to be discussed ahead) if there are two or more explanatory variables.
Simple Linear Regression

- It is possible to show using calculus that the values of $b_0$ and $b_1$ minimising the sum of squared vertical deviations of the sample pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ are given by:

  - $b_0 = \bar{Y} - b_1 \bar{X}$
  - $b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$

- This is known as the **method of least squares**.
Simple Linear Regression

- The use of $b_0$ and $b_1$ rather than $\beta_0$ and $\beta_1$ indicates that $b_0$ and $b_1$ are sample estimates of the true values: a different sample would give different estimates.
The use of $b_0$ and $b_1$ rather than $\beta_0$ and $\beta_1$ indicates that $b_0$ and $b_1$ are sample estimates of the true values: a different sample would give different estimates.

The slope coefficient estimate $b_1$ is related to the sample correlation coefficient $r_{XY}$ as follows:

$$b_1 = r_{XY} \frac{s_Y}{s_X} = r_{XY} \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$
Simple Linear Regression

- The use of $b_0$ and $b_1$ rather than $\beta_0$ and $\beta_1$ indicates that $b_0$ and $b_1$ are sample estimates of the true values: a different sample would give different estimates.

- The slope coefficient estimate $b_1$ is related to the sample correlation coefficient $r_{XY}$ as follows:

$$b_1 = r_{XY} \frac{s_Y}{s_X} = r_{XY} \frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- Because $s_X$ and $s_Y$ are non-negative, $b_1$ will have the same sign as $r_{XY}$: the regression line will slope upwards if the observed linear association is positive and downwards if it is negative.
Example 12: The following table gives data collected last year for seven employees of a company, where $X=$ number of years of service and $Y=$ number of days taken off work.

\[
\begin{array}{cccccccc}
X & 2 & 5 & 7 & 3 & 8 & 3 & 7 \\
Y & 8 & 7 & 5 & 12 & 3 & 9 & 5 \\
\end{array}
\]
Example 12: The following table gives data collected last year for seven employees of a company, where $X =$ number of years of service and $Y =$ number of days taken off work.

<table>
<thead>
<tr>
<th>$X$</th>
<th>2</th>
<th>5</th>
<th>7</th>
<th>3</th>
<th>8</th>
<th>3</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>8</td>
<td>7</td>
<td>5</td>
<td>12</td>
<td>3</td>
<td>9</td>
<td>5</td>
</tr>
</tbody>
</table>

Hence we have $\bar{X} = \frac{35}{7} = 5$, $\bar{Y} = \frac{49}{7} = 7$ and

\[
(X - \bar{X})^2 = 9, 0, 4, 4, 9, 4, 4 \quad \sum = 34
\]
\[
(Y - \bar{Y})^2 = 1, 0, -2, 5, -4, 2, -2 \quad \sum = 54
\]
\[
(X - \bar{X})(Y - \bar{Y}) = -3, 0, -4, -10, -12, -4, -4 \quad \sum = -37
\]
Therefore,
\[ r_{XY} = -\frac{37}{\sqrt{(34)(54)}} = -0.864, \]
\[ b_1 = -\frac{37}{34} = -1.09, \text{ or } b_1 = -0.864 \frac{\sqrt{54}}{\sqrt{34}} = -1.09 \]
\[ b_0 = 7 - (-1.09)5 = 12.45, \text{ and} \]
Therefore, 
\[ r_{XY} = -\frac{37}{\sqrt{(34)(54)}} = -0.864, \]
\[ b_1 = -\frac{37}{34} = -1.09, \]
\[ b_0 = 7 - (-1.09)5 = 12.45, \]
and the estimated regression line is \( \hat{Y} = 12.45 - 1.09X \).
Therefore,

\[ r_{XY} = -\frac{37}{\sqrt{(34)(54)}} = -0.864, \]

\[ b_1 = -\frac{37}{34} = -1.09, \text{ or } b_1 = -0.864 \frac{\sqrt{54}}{\sqrt{34}} = -1.09 \]

\[ b_0 = 7 - (-1.09)5 = 12.45, \text{ and } \]

the estimated regression line is \( \hat{Y} = 12.45 - 1.09X. \)

The "hat" on \( Y \) represents prediction; \( \hat{Y} \) is the predicted or forecasted value of \( Y \) for a given value of \( X \). Clearly, \( \hat{Y} \neq Y. \hat{Y} = Y \) for all sample values if and only if \( |r_{XY}| = 1. \)
Simple Linear Regression

![Simple Linear Regression Graph](image_url)
Simple Linear Regression

- The analysis can be easily performed by EXCEL with output as follows (to be discussed in tutorials):
Simple Linear Regression

The analysis can be easily performed by EXCEL with output as follows (to be discussed in tutorials):

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.863506052</td>
</tr>
<tr>
<td>R Square</td>
<td>0.745642702</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.694771242</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.65742536</td>
</tr>
<tr>
<td>Observations</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.44117647</td>
<td>1.553168751</td>
<td>8.01019</td>
<td>0.00049</td>
<td>8.44862909</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>1.088235294</td>
<td>0.284246104</td>
<td>-3.8285</td>
<td>0.012266</td>
<td>1.81891317</td>
</tr>
</tbody>
</table>
Simple Linear Regression

- The analysis can be easily performed by EXCEL with output as follows (to be discussed in tutorials):

```
SUMMARY OUTPUT

Regression Statistics

Multiple R 0.863506052
R Square 0.745642702
Adjusted R Square 0.694771242
Standard Error 1.65742536
Observations 7

Coefficients Standard Error t Stat P-value Lower 95% Upper 95%
Intercept 12.44117647 1.553168751 8.01019 0.00049 8.44862909 16.43372385
X Variable 1 1.088235294 0.284246104 -3.8285 0.012266 1.81891317 -0.357557422
```
Simple Linear Regression

Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

\[
\hat{Y} = 12.45 - 1.09X
\]
Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

$\hat{Y} = 12.45 - 1.09(5) = 7$ days off work
Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

$\hat{Y} = 12.45 - 1.09(5) = 7$ days off work

$\hat{Y} = 12.45 - 1.09(6) = 5.91$ (i.e., about 6) days off work
Simple Linear Regression

Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

$\hat{Y} = 12.45 - 1.09(X) = \text{number of days off work}$

- $\hat{Y} = 12.45 - 1.09(5) = 7$ days off work
- $\hat{Y} = 12.45 - 1.09(6) = 5.91$ (i.e., about 6) days off work
- $\hat{Y} = 12.45 - 1.09(8) = 3.73$ (i.e., about 4) days off work
- $\hat{Y} = 12.45 - 1.09(0) = 12.45$ (i.e., about 12) days off work
- $\hat{Y} = 12.45 - 1.09(14) = -2.81$ days off work
Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

$\hat{Y} = 12.45 - 1.09X$

- $\hat{Y} = 12.45 - 1.09(5) = 7$ days off work
- $\hat{Y} = 12.45 - 1.09(6) = 5.91$ (i.e., about 6) days off work
- $\hat{Y} = 12.45 - 1.09(8) = 3.73$ (i.e., about 4) days off work
- $\hat{Y} = 12.45 - 1.09(0) = 12.45$ (i.e., about 12) days off work
Simple Linear Regression

Suppose we want to predict the number of days off work this year for employees with 5, 6, 8, 0 and 14 years of service. All we have to do is to substitute the given $X$ values into the estimated regression equation:

$\hat{Y} = 12.45 - 1.09(5) = 7$ days off work

$\hat{Y} = 12.45 - 1.09(6) = 5.91$ (i.e., about 6) days off work

$\hat{Y} = 12.45 - 1.09(8) = 3.73$ (i.e., about 4) days off work

$\hat{Y} = 12.45 - 1.09(0) = 12.45$ (i.e., about 12) days off work

$\hat{Y} = 12.45 - 1.09(14) = -2.81$ days off work (What ????)
Interpreting $b_0$: Note that from the prediction of $Y$ for $X = 0$, we see that $b_0 = 12.45$ is the predicted number of days off for an employee with 0 years of service (i.e., a new employee).
Simple Linear Regression

- Interpreting $b_0$: Note that from the prediction of $Y$ for $X = 0$, we see that $b_0 = 12.45$ is the predicted number of days off for an employee with 0 years of service (i.e., a new employee).

- Interpreting $b_1$: Subtracting the prediction for $X = 5$ (i.e., $\hat{Y} = 7$) from the prediction for $X = 6$ (i.e., $\hat{Y} = 5.91$) gives $b_1 = -1.09$: thus $b_1$ is the change in the estimated number of days off for an additional year’s service.
3.1 Simple Linear Regression

Simple Linear Regression

- Interpreting $b_0$: Note that from the prediction of $Y$ for $X = 0$, we see that $b_0 = 12.45$ is the predicted number of days off for an employee with 0 years of service (i.e., a new employee).

- Interpreting $b_1$: Subtracting the prediction for $X = 5$ (i.e., $\hat{Y} = 7$) from the prediction for $X = 6$ (i.e., $\hat{Y} = 5.91$) gives $b_1 = -1.09$: thus $b_1$ is the change in the estimated number of days off for an additional year’s service.

- In general, $b_0$ is the predicted value of $Y$ for $X = 0$, while $b_1$ is the marginal change in $Y$ for a one unit’s increase in $X$. 

Finally, note that the prediction of -2.81 days off work for \( X=14 \) years of service does not make sense: what meaning can we attach to a negative number of days off work?
Finally, note that the prediction of -2.81 days off work for \( X = 14 \) years of service does not make sense: what meaning can we attach to a negative number of days off work?

The scatter diagram may suggest the reason for this peculiar value: the relationship between \( X \) and \( Y \) is approximately linear over the range covered by the sample, but the regression line cannot be extended indefinitely without cutting the \( X \)-axis. Once we go beyond the sample range the relationship may cease to be approximately linear.
Simple Linear Regression

Example 13: The file HKpopulation.xlsx contains yearly data on Hong Kong’s population from 1960 to 2013, totalling 54 observations. Let $Y$ denote the population figure and $X = 1, 2, 3, \ldots$ denote the sequence of time, with $X = 1$ representing the year 1960, $X = 2$ representing 1961, etc.
Simple Linear Regression

- Example 13: The file HKpopulation.xlsx contains yearly data on Hong Kong’s population from 1960 to 2013, totalling 54 observations. Let $Y$ denote the population figure and $X = 1, 2, 3,...$ denote the sequence of time, with $X = 1$ representing the year 1960, $X = 2$ representing 1961, etc.

- An excerpt of the data is as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1960</td>
<td>1</td>
<td>2619983</td>
</tr>
<tr>
<td>1961</td>
<td>2</td>
<td>2735407</td>
</tr>
<tr>
<td>1969</td>
<td>10</td>
<td>3402250</td>
</tr>
<tr>
<td>1998</td>
<td>39</td>
<td>6510390</td>
</tr>
<tr>
<td>2013</td>
<td>54</td>
<td>7259603</td>
</tr>
</tbody>
</table>
A plot of $Y$ versus $X$ reveals the following:

![Graph showing population over time with a trend line indicating a positive correlation.](image-url)
Simple Linear Regression

- A plot of $Y$ versus $X$ reveals the following:

- The association between $X$ and $Y$ appears to be approximately linear. It therefore makes sense to write $Y = \beta_0 + \beta_1 X + u$
Simple Linear Regression

- Using the aforementioned least squares method, we can compute the following estimates $b_0$ and $b_1$ (the computation can be done automatically in EXCEL):
  
  $b_0 = 2597679.706$ and $b_1 = 93778.51912$
Simple Linear Regression

Using the aforementioned least squares method, we can compute the following estimates $b_0$ and $b_1$ (the computation can be done automatically in EXCEL):

\[ b_0 = 2597679.706 \quad \text{and} \quad b_1 = 93778.51912 \]

**SUMMARY OUTPUT**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2597679.706</td>
<td>49287.04567</td>
<td>52.70512</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>93778.51912</td>
<td>1559.243055</td>
<td>60.14362</td>
</tr>
</tbody>
</table>
Simple Linear Regression

Using the aforementioned least squares method, we can compute the following estimates $b_0$ and $b_1$ (the computation can be done automatically in EXCEL):

$b_0 = 2597679.706$ and $b_1 = 93778.51912$

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2597679.706</td>
<td>49287.04567</td>
<td>52.70512</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>93778.51912</td>
<td>1559.243055</td>
<td>60.14362</td>
</tr>
</tbody>
</table>
So, $b_0 = 2597679.706$ is the predicted Hong Kong population for the year 1959 ($X = 0$) and $b_1 = 93778.51912$ is the predicted average annual increment in population.
So, $b_0 = 2597679.706$ is the predicted Hong Kong population for the year 1959 ($X = 0$) and $b_1 = 93778.51912$ is the predicted average annual increment in population.

Now, the in-sample predictions for the years 1998 ($X = 39$) and 2013 ($X = 54$) are:

$\hat{Y}_{39} = 2597679.706 + 93778.51912(39) = 6255042$, and $\hat{Y}_{54} = 2597679.706 + 93778.51912(54) = 7661720$ respectively.
Simple Linear Regression

But \( Y_{39} = 6510390 \) and \( Y_{54} = 7259603 \), so the forecast errors (denoted by \( e = Y - \hat{Y} \)) are:

\[
e_{39} = 6510390 - 6255042 = 255348 \quad \text{and} \quad e_{54} = 7259603 - 7661720 = -402116
\]
But $Y_{39} = 6510390$ and $Y_{54} = 7259603$, so the forecast errors (denoted by $e = Y - \hat{Y}$) are:

\[ e_{39} = 6510390 - 6255042 = 255348 \quad \text{and} \quad e_{54} = 7259603 - 7661720 = -402116 \]

So, our model ”under-estimates” the population in 1989 but ”over-estimates” the population in 2013. This does not mean our model is bad as the regression can never make a precise prediction without errors unless the linear association is perfect.
Simple Linear Regression

- But $Y_{39} = 6510390$ and $Y_{54} = 7259603$, so the forecast errors (denoted by $e = Y - \hat{Y}$) are:
  
  $e_{39} = 6510390 - 6255042 = 255348$ and $e_{54} = 7259603 - 7661720 = -402116$

- So, our model "under-estimates" the population in 1989 but "over-estimates" the population in 2013. This does not mean our model is bad as the regression can never make a precise prediction without errors unless the linear association is perfect.

- In fact, the "$R^2$" from the regression output has a value of 0.9858, indicating that the estimated regression has the ability to capture 98.58% of the variation in $Y$ in the sample.
Simple Linear Regression

$R^2$ is known as the coefficient of determination. It measures the goodness of fit of the regression model in terms of the proportion of variation in $Y$ explained by the regression, and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$
Simple Linear Regression

- $R^2$ is known as the coefficient of determination. It measures the goodness of fit of the regression model in terms of the proportion of variation in $Y$ explained by the regression, and is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

- It can be shown that $0 \leq R^2 \leq 1$. Values of $R^2$ close to 1 indicate that the model fits the data well. Conversely, values of $R^2$ close to 0 indicate poor fit.
Simple Linear Regression

Thus, in our population forecast example, the regression model accounts for 98.58% of the variation in $Y$; the remaining 1.42% of the variation that cannot be captured by the regression is reflected in the error term $e$. 
Thus, in our population forecast example, the regression model accounts for 98.58% of the variation in $Y$; the remaining 1.42% of the variation that cannot be captured by the regression is reflected in the error term $e$.

It can be shown that $\sum_{i=1}^{n} e_i$ is always zero, so the total amounts of over- and under-estimation across the sample values always offset each other.
Simple Linear Regression

▶ Thus, in our population forecast example, the regression model accounts for 98.58% of the variation in $Y$; the remaining 1.42% of the variation that cannot be captured by the regression is reflected in the error term $e$.

▶ It can be shown that $\sum_{i=1}^{n} e_i$ is always zero, so the total amounts of over- and under-estimation across the sample values always offset each other.

▶ In a regression model containing only one $X$ variable, $R^2 = r_{XY}^2$. Hence in our population forecast example, the sample correlation coefficient between $X$ and $Y$ is $r_{XY} = +\sqrt{0.9858} = 0.9929$. We know $r_{XY}$ has a positive sign because $b_1$ is positive; $r_{XY}$ would have a negative sign if $b_1$ were negative.
Suppose we want to forecast the future Hong Kong population for 2014 - 2020. Note that 2014 corresponds to $X = 55$ and so on. The forecasts are therefore:

2014: $\hat{Y}_{55} = 2597679.706 + 93778.51912(55) = 7755498$
2015: $\hat{Y}_{56} = 2597679.706 + 93778.51912(56) = 7849277$
2016: $\hat{Y}_{57} = 2597679.706 + 93778.51912(57) = 7943055$
2017: $\hat{Y}_{58} = 2597679.706 + 93778.51912(58) = 8036834$
2018: $\hat{Y}_{59} = 2597679.706 + 93778.51912(59) = 8130612$
2019: $\hat{Y}_{60} = 2597679.706 + 93778.51912(60) = 8224391$
2020: $\hat{Y}_{61} = 2597679.706 + 93778.51912(61) = 8318169$
Simple Linear Regression

- Of course, the accuracy of these forecasts depends, among other things, on the legitimacy to extend the linear relationship established based on the sample values beyond the estimation period.

- These forecasts are called "ex-ante" forecasts since the actual values of the variable being predicted are unknown at the time of prediction.
Simple Linear Regression

In many studies, the primary objective of regression is to estimate the slope coefficient $\beta_1$, which indicates the change in $Y$ when $X$ changes by one unit. In our first example, $\beta_1$ is the change in the number of days off for an additional year’s service. Note that $\beta_1$ is the actual change which can never be known. It is important to distinguish between $\beta_1$ and $b_1$: $b_1$ is merely an estimate of $\beta_1$ and not the same as $\beta_1$. 

Note that if $\beta_1 = 0$, then there is really no relationship between $Y$ when $X$; that is, a change in $X$ does not induce any change in $Y$. That said, even when $\beta_1 = 0$, $b_1$ is almost certainly non-zero as it is merely an estimate of $\beta_1$. 

55 / 74
In many studies, the primary objective of regression is to estimate the slope coefficient $\beta_1$, which indicates the change in $Y$ when $X$ changes by one unit. In our first example, $\beta_1$ is the change in the number of days off for an additional year’s service. Note that $\beta_1$ is the actual change which can never be known. It is important to distinguish between $\beta_1$ and $b_1$: $b_1$ is merely an estimate of $\beta_1$ and not the same as $\beta_1$.

Note that if $\beta_1 = 0$, then there is really no relationship between $Y$ when $X$; that is, a change in $X$ does not induce any change in $Y$. That said, even when $\beta_1 = 0$, $b_1$ is almost certainly non-zero as it is merely an estimate of $\beta_1$. 
Simple Linear Regression

A test whether a relationship between $X$ and $Y$ exists is therefore appropriate. The test takes the form:

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$
Simple Linear Regression

▶ A test whether a relationship between $X$ and $Y$ exists is therefore appropriate. The test takes the form:

$H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$

▶ The test statistic is

$t = \frac{b_1 - 0}{S_{b_1}} \sim t_{n-2}$, where $S_{b_1}$ is the standard error of $b_1$ that measures the variation of $b_1$. 

A test whether a relationship between $X$ and $Y$ exists is therefore appropriate. The test takes the form:

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

The test statistic is

$$t = \frac{b_1 - 0}{S_{b_1}} \sim t_{n-2},$$

where $S_{b_1}$ is the standard error of $b_1$ that measures the variation of $b_1$.

At $\alpha = 0.05$, we reject $H_0$ if the p-value of the test is smaller than 0.05, or $|t| > t(\alpha/2, n-2)$. 
Simple Linear Regression

- Rejecting $H_0$ is the more desirable outcome as it suggests $eta_1 \neq 0$, meaning that $X$ does have an impact on the behaviour of $Y$, and we have chosen a correct explanatory variable.
Simple Linear Regression

- Rejecting $H_0$ is the more desirable outcome as it suggests $\beta_1 \neq 0$, meaning that $X$ does have an impact on the behaviour of $Y$, and we have chosen a correct explanatory variable.

- If $H_0$ is rejected, we say that $\beta_1$ significantly differs from zero or $X$ is significant.
Simple Linear Regression

- Rejecting $H_0$ is the more desirable outcome as it suggests $\beta_1 \neq 0$, meaning that $X$ does have an impact on the behaviour of $Y$, and we have chosen a correct explanatory variable.
- If $H_0$ is rejected, we say that $\beta_1$ significantly differs from zero or $X$ is significant.
- Alternatively, if $H_0$ cannot be rejected, we say that $\beta_1$ does not differ from zero significantly or $X$ is insignificant.
Example 14: Consider the data and the regression of Example 12. Suppose we want to test if years of service is linearly associated with the number of days off work.

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \]
Simple Linear Regression

- Example 14: Consider the data and the regression of Example 12. Suppose we want to test if years of service is linearly associated with the number of days off work.

- $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

- $\alpha = 0.05$, $n = 7$, $df = 5$ and $t(0.05/2, 5) = 2.5706$ (from table).

- $t = b_1 - 0 \frac{S_{b_1}}{0.2842} = -3.835$

- As $|−3.835| > 2.5706$ or $p-value < 0.05$, we REJECT $H_0$ at $\alpha = 0.05$ and conclude that years of service is linearly related to the number of days off work.
Example 14: Consider the data and the regression of Example 12. Suppose we want to test if years of service is linearly associated with the number of days off work.

- \( H_0 : \beta_1 = 0 \) vs. \( H_1 : \beta_1 \neq 0 \)
- \( \alpha = 0.05 \), \( n = 7 \), \( df = 5 \) and \( t_{(0.05/2, 5)} = 2.5706 \) (from table).
- \( t = \frac{b_1 - 0}{S_{b_1}} = \frac{-1.09 - 0}{0.2842} = -3.835 \)

As \(|-3.835| > 2.5706\) or \(p\)-value < 0.05, we REJECT \( H_0 \) at \( \alpha = 0.05 \) and conclude that years of service is linearly related to the number of days off work.
Example 14: Consider the data and the regression of Example 12. Suppose we want to test if years of service is linearly associated with the number of days off work.

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \]

\[ \alpha = 0.05, \quad n = 7, \quad df = 5 \quad \text{and} \quad t_{(0.05/2, \ 5)} = 2.5706 \quad \text{(from table)} \]

\[ t = \frac{b_1 - 0}{S_{b_1}} = \frac{-1.09 - 0}{0.2842} = -3.835 \]

As \(| -3.835 | > 2.5706\) or p-value < 0.05, we REJECT \(H_0\) at \(\alpha = 0.05\) and conclude that years of service is linearly related to the number of days off work.
Simple Linear Regression

- This test outcome is unsurprising as $R^2 = 0.7456$ is high.
Simple Linear Regression

- This test outcome is unsurprising as $R^2 = 0.7456$ is high.
- Testing $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ is equivalent to asking if the 95% confidence interval of $\beta_1$ contains 0.

\[ b_1 \pm t(0.05/2, n-2) \times S_{b_1} = -1.09 \pm 2.5706 \times 0.2842 = [-1.821, -0.359] \]

This CI is consistent with the test outcome. Given we reject $H_0 : \beta_1 = 0$, we do not expect the 95% confidence interval of $\beta_1$ to contain 0.
Simple Linear Regression

- This test outcome is unsurprising as $R^2 = 0.7456$ is high.
- Testing $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ is equivalent to asking if the 95% confidence interval of $\beta_1$ contains 0.
- The 95% confidence interval of $\beta_1$ can be found by

$$b_1 \pm t_{(0.05/2,n-2)} S_{b_1} = -1.09 \pm 2.5706 \times 0.2842$$

$$= [-1.82, -0.359]$$
This test outcome is unsurprising as $R^2 = 0.7456$ is high.

Testing $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ is equivalent to asking if the 95% confidence interval of $\beta_1$ contains 0.

The 95% confidence interval of $\beta_1$ can be found by

$$b_1 \pm t_{(0.05/2, n-2)} S_{b_1} = -1.09 \pm 2.5706 \times 0.2842$$

$$= [-1.821, -0.359]$$

This CI is consistent with the test outcome. Given we reject $H_0 : \beta_1 = 0$, we do not expect the 95% confidence interval of $\beta_1$ to contain 0.
Simple Linear Regression

- One can conduct hypothesis test and construct CI in the same manner for $\beta_0$, but usually inferences on the intercept are of less interest.
- Also, $R^2$ loses its goodness of fit meaning if we remove the intercept from the model.
Simple Linear Regression

- One can conduct hypothesis test and construct CI in the same manner for $\beta_0$, but usually inferences on the intercept are of less interest.
- Also, $R^2$ loses its goodness of fit meaning if we remove the intercept from the model.
- The usual practice is to always include the intercept even if it is insignificant.
Multiple Linear Regression

- In many situations, two or more explanatory variables may be included in a regression model to provide an adequate description of the process under study or to yield sufficiently precise inferences.
Multiple Linear Regression

- In many situations, two or more explanatory variables may be included in a regression model to provide an adequate description of the process under study or to yield sufficiently precise inferences.

- For example, a regression model for predicting the demands for a firm’s product in different countries uses socioeconomic variables (mean household income, average years of schooling of head of household), demographic variables (average family size, percentage of retired population), and environmental variables (mean daily temperature, pollution index), etc.
Multiple Linear Regression

- In many situations, two or more explanatory variables may be included in a regression model to provide an adequate description of the process under study or to yield sufficiently precise inferences.

- For example, a regression model for predicting the demands for a firm’s product in different countries uses socioeconomic variables (mean household income, average years of schooling of head of household), demographic variables (average family size, percentage of retired population), and environmental variables (mean daily temperature, pollution index), etc.

- Linear regression models containing two or more explanatory variables are called **multiple linear regression** models.
Multiple Linear Regression

The simple linear regression model discussed earlier can be extended to include more than one explanatory variable:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u, \]

where \( u \) is an unobserved error term and \( \beta_0, \beta_1,\ldots, \beta_k \) are unknown coefficients to be estimated.
The simple linear regression model discussed earlier can be extended to include more than one explanatory variables:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u, \]

where \( u \) is an unobserved error term and \( \beta_0, \beta_1, ..., \beta_k \) are unknown coefficients to be estimated.

There are \( k \) explanatory variables, namely, \( X_1, X_2, ... X_k \).
The simple linear regression model discussed earlier can be extended to include more than one explanatory variables:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u, \]

where \( u \) is an unobserved error term and \( \beta_0, \beta_1, ..., \beta_k \) are unknown coefficients to be estimated.

- There are \( k \) explanatory variables, namely, \( X_1, X_2, ..., X_k \).
- When \( k = 1 \), the multiple linear regression model reduces to the simple linear regression model.
The simple linear regression model discussed earlier can be extended to include more than one explanatory variables:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u, \]

where \( u \) is an unobserved error term and \( \beta_0, \beta_1, \ldots \) and \( \beta_k \) are unknown coefficients to be estimated.

There are \( k \) explanatory variables, namely, \( X_1, X_2, \ldots X_k \).

When \( k = 1 \), the multiple linear regression model reduces to the simple linear regression model.
Multiple Linear Regression

As in the case of simple linear regression, we use the method of least squares to estimate $\beta_0$, $\beta_1$,..., and $\beta_k$. The estimates are labelled as $b_0$, $b_1$,..., and $b_k$. 
As in the case of simple linear regression, we use the method of least squares to estimate $\beta_0$, $\beta_1$, ..., and $\beta_k$. The estimates are labelled as $b_0$, $b_1$, ..., and $b_k$.

The formulae for $b_0$, $b_1$, ..., and $b_k$ are complicated, and their complexity increases as $k$ increases. Inferences under multiple linear regression are also more complex.
Multiple Linear Regression

- As in the case of simple linear regression, we use the method of least squares to estimate $\beta_0$, $\beta_1$, ..., and $\beta_k$. The estimates are labelled as $b_0$, $b_1$, ..., and $b_k$.
- The formulae for $b_0$, $b_1$, ..., and $b_k$ are complicated, and their complexity increases as $k$ increases. Inferences under multiple linear regression are also more complex.
- We will *not* be concerned about the manual computation of $b_0$, $b_1$, ..., and $b_k$ here - we will simply use EXCEL to compute these estimates.
As in the case of simple linear regression, we use the method of least squares to estimate $\beta_0$, $\beta_1$, ..., and $\beta_k$. The estimates are labelled as $b_0$, $b_1$, ..., and $b_k$.

The formulae for $b_0$, $b_1$, ..., and $b_k$ are complicated, and their complexity increases as $k$ increases. Inferences under multiple linear regression are also more complex.

We will not be concerned about the manual computation of $b_0$, $b_1$, ..., and $b_k$ here - we will simply use EXCEL to compute these estimates.

Our discussion will only "touch on" multiple regression through a case study. The technical details of multiple regression will be covered in details in more advanced level courses.
Example 15: A case study on multiple regression

The Grade Point Average (GPA) is commonly adopted by universities as the indication of students’ academic performance. There is one common consensus that the Intelligence Quotient (IQ) is a significant factor affecting the GPA of a student. On the other hand, it has been suggested that academic results are also related to the student’s lifestyle. In 2012, the Student Development Services at City U conducted a statistical study to examine the relationship between GPA and a number of potential factors.
Multiple Linear Regression

One hundred and twenty 1st-year Business students took part in this study. These participants were selected randomly. Each took an IQ test and had to answer questions relating to his/her study and lifestyle. The data are contained in the EXCEL file Students.xlsx. They are described briefly as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>Explanatory Variables (X)</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligence</td>
<td>IQ</td>
<td>Result in the Intelligence Quotient test</td>
<td>110 – 138</td>
</tr>
<tr>
<td>Study load</td>
<td>NoCrs</td>
<td>Number of credits taken in the semester A</td>
<td>12, 15, 18</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>TravelMin</td>
<td>Average minutes required to travel between home and CityU</td>
<td>0, 20, 40, 60, 80, 100, 120</td>
</tr>
<tr>
<td></td>
<td>SleepHr</td>
<td>Average hours spent on sleeping per day</td>
<td>6, 8, 10, 12</td>
</tr>
<tr>
<td></td>
<td>StudyHr</td>
<td>Average hours spent on studying per week</td>
<td>0, 5, 10, 15, 20, 25, 30, 35, 40</td>
</tr>
<tr>
<td></td>
<td>WorkHr</td>
<td>Average hours spent on working per week</td>
<td>0, 5, 10, 15, 20, 25, 30, 35, 40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Dependent Variable (Y)</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic achievement</td>
<td>SGPA</td>
<td>Grade Point Average in the semester A</td>
<td>1.29 – 4.30</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

One hundred and twenty 1st-year Business students took part in this study. These participants were selected randomly. Each took an IQ test and had to answer questions relating to his/her study and lifestyle. The data are contained in the EXCEL file Students.xlsx. They are described briefly as follows:

<table>
<thead>
<tr>
<th>Category</th>
<th>Explanatory Variables (X)</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intelligence</td>
<td>IQ</td>
<td>Result in the Intelligence Quotient test</td>
<td>110 – 138</td>
</tr>
<tr>
<td>Study load</td>
<td>NoCrs</td>
<td>Number of credits taken in the semester A</td>
<td>12, 15, 18</td>
</tr>
<tr>
<td>Lifestyle</td>
<td>TravelMin</td>
<td>Average minutes required to travel between home and CityU</td>
<td>0, 20, 40, 60, 80, 100, 120</td>
</tr>
<tr>
<td></td>
<td>SleepHr</td>
<td>Average hours spent on sleeping per day</td>
<td>6, 8, 10, 12</td>
</tr>
<tr>
<td></td>
<td>StudyHr</td>
<td>Average hours spent on studying per week</td>
<td>0, 5, 10, 15, 20, 25, 30, 35, 40</td>
</tr>
<tr>
<td></td>
<td>WorkHr</td>
<td>Average hours spent on working per week</td>
<td>0, 5, 10, 15, 20, 25, 30, 35, 40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Dependent Variable (Y)</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic achievement</td>
<td>SGPA</td>
<td>Grade Point Average in the semester A</td>
<td>1.29 – 4.30</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

- The multiple regression model is

\[ \text{SGPA} = \beta_0 + \beta_1 IQ + \beta_2 \text{NoCrs} + \beta_3 \text{TravelMin} + \beta_4 \text{SleepHr} + \beta_5 \text{StudyHr} + \beta_6 \text{WorkHr} + u. \]

- Note that not all of the six explanatory variables are necessarily significant, meaning that the performance of the regression may be ”better off” with some of the explanatory variables on the right-hand-side of the equation removed.
The multiple regression model is

\[
SGPA = \beta_0 + \beta_1 IQ + \beta_2 NoCrs + \beta_3 TravelMin \\
+ \beta_4 SleepHr + \beta_5 StudyHr + \beta_6 WorkHr + u.
\]

Note that not all of the six explanatory variables are necessarily significant, meaning that the performance of the regression may be ”better off” with some of the explanatory variables on the right-hand-side of the equation removed.

We can test for the significance of each explanatory variable using a t-test after estimating the above full model. The results are shown as follows.
Multiple Linear Regression

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.913462442</td>
<td>0.680455298</td>
<td>1.34242829</td>
<td>0.182147891</td>
<td>-0.434642229</td>
</tr>
<tr>
<td>IQ</td>
<td>0.022195038</td>
<td>0.004903885</td>
<td>4.52601053</td>
<td>1.49509E-05</td>
<td>0.012479557</td>
</tr>
<tr>
<td>NoCrs</td>
<td>-0.048683053</td>
<td>0.024812642</td>
<td>-1.962026154</td>
<td>0.052218261</td>
<td>-0.097841372</td>
</tr>
<tr>
<td>TravelMin</td>
<td>0.000848208</td>
<td>0.000796475</td>
<td>1.064952445</td>
<td>0.289167731</td>
<td>-0.000729753</td>
</tr>
<tr>
<td>SleepHr</td>
<td>0.012080628</td>
<td>0.023950257</td>
<td>0.504404941</td>
<td>0.614958993</td>
<td>-0.035369151</td>
</tr>
<tr>
<td>StudyHr</td>
<td>0.031439423</td>
<td>0.004013515</td>
<td>7.83388222</td>
<td>2.82763E-12</td>
<td>0.023487926</td>
</tr>
<tr>
<td>WorkHr</td>
<td>-0.02982413</td>
<td>0.003813774</td>
<td>-7.82011011</td>
<td>3.02793E-12</td>
<td>-0.037379903</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

- $R^2$ is 0.81599, meaning that the model accounts for nearly 82% of the variation of the dependent variable, $SGPA$, in the sample.
R² is 0.81599, meaning that the model accounts for nearly 82% of the variation of the dependent variable, SGPA, in the sample.

A test of whether a relationship exists between the dependent variable SGPA and a given explanatory variable takes the form:

\[ H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0 \]

The test statistic is

\[ t = \frac{b_i - 0}{S_{b_i}} \sim t_{n-(k+1)}, \text{ where } S_{b_i} \text{ is the standard error of } b_i, \]

and \( i \) is an index representing one of 1, 2, 3, .... \( k \).
★ $R^2$ is 0.81599, meaning that the model accounts for nearly 82% of the variation of the dependent variable, $SGPA$, in the sample.

★ A test of whether a relationship exists between the dependent variable $SGPA$ and a given explanatory variable takes the form:

$$H_0 : \beta_i = 0 \quad \text{vs.} \quad H_1 : \beta_i \neq 0$$

★ The test statistic is

$$t = \frac{b_i - 0}{S_{b_i}} \sim t_{n-(k+1)}$$

where $S_{b_i}$ is the standard error of $b_i$, and $i$ is an index representing one of 1, 2, 3, ..., $k$.

★ At $\alpha = 0.05$, we reject $H_0$ if the p-value of the test is smaller than 0.05, or $|t| > t(\alpha/2, n-(k+1))$. 
Multiple Linear Regression

- $R^2$ is 0.81599, meaning that the model accounts for nearly 82% of the variation of the dependent variable, $SGPA$, in the sample.
- A test of whether a relationship exists between the dependent variable $SGPA$ and a given explanatory variable takes the form: $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$
- The test statistic is $t = \frac{b_i - 0}{S_{b_i}} \sim t_{n-(k+1)}$, where $S_{b_i}$ is the standard error of $b_i$, and $i$ is an index representing one of $1, 2, 3, \ldots, k$.
- At $\alpha = 0.05$, we reject $H_0$ if the p-value of the test is smaller than 0.05, or $|t| > t(\alpha/2, n-(k+1))$.
- If $H_0$ cannot be rejected, there is a good chance that $\beta_i$ is zero and the regression should be re-estimated with the explanatory variable associated with $\beta_i$ removed.
One can, in addition to $R^2$, evaluate the model’s goodness of fit by the sum of the absolute values of errors across the sample of observations.
One can, in addition to $R^2$, evaluate the model’s goodness of fit by the sum of the absolute values of errors across the sample of observations.

Recall that the error for observation $i$ is $e_i = Y_i - \hat{Y}_i$, or using the notations of the current example, $e_i = SGPA_i - \hat{SGPA}_i$. It measures the extent to which the predicted value of the dependent variable differs from its true value for an observation.
One can, in addition to $R^2$, evaluate the model’s goodness of fit by the sum of the absolute values of errors across the sample of observations.

Recall that the error for observation $i$ is $e_i = Y_i - \hat{Y}_i$, or using the notations of the current example, $e_i = SGPA_i - \hat{SGPA}_i$. It measures the extent to which the predicted value of the dependent variable differs from its true value for an observation.

The sum of the absolute values of errors (SAE) is $\sum_i^n |e_i|$. We consider the sum of $|e_i|$ rather than the sum of $e_i$ to avoid negative $e_i$’s cancelling out with positive $e_i$’s.
Multiple Linear Regression

- One can, in addition to $R^2$, evaluate the model’s goodness of fit by the sum of the absolute values of errors across the sample of observations.

- Recall that the error for observation $i$ is $e_i = Y_i - \hat{Y}_i$, or using the notations of the current example, $e_i = SGPA_i - \hat{SGPA}_i$. It measures the extent to which the predicted value of the dependent variable differs from its true value for an observation.

- The sum of the absolute values of errors (SAE) is $\sum_{i}^{n} |e_i|$. We consider the sum of $|e_i|$ rather than the sum of $e_i$ to avoid negative $e_i$’s cancelling out with positive $e_i$’s.

- Clearly, the smaller the SAE, the better the fit of the model.
Multiple Linear Regression

- For example, using the model and the data in "students.xlsx", we obtain

\[
\hat{SGPA}_1 = 0.913462442 + 0.022195038(116) - 0.048683053(15) \\
+ 0.000848208(60) + 0.012080628(12) + 0.031439423(25) \\
- 0.02982413(5) = 3.5905662,
\]

which is the predicted value of \( SGPA_1 \).
Multiple Linear Regression

For example, using the model and the data in "students.xlsx", we obtain

\[
\hat{SGPA}_1 = 0.913462442 + 0.022195038(116) - 0.048683053(15) \\
+ 0.000848208(60) + 0.012080628(12) + 0.031439423(25) \\
- 0.02982413(5) = 3.5905662,
\]

which is the predicted value of \( SGPA_1 \).

From the data provided, the true value of \( SGPA_1 \) is 3.98. Hence \( e_1 = 3.98 - 3.5905662 = 0.38943398 \) and \( |e_1| = 0.38943398 \).
Multiple Linear Regression

- For example, using the model and the data in "students.xlsx", we obtain

\[
\hat{SGPA}_1 = 0.913462442 + 0.022195038(116) - 0.048683053(15) \\
+ 0.000848208(60) + 0.012080628(12) + 0.031439423(25) \\
- 0.02982413(5) = 3.5905662,
\]

which is the predicted value of \( SGPA_1 \).

- From the data provided, the true value of \( SGPA_1 \) is 3.98. Hence \( e_1 = 3.98 - 3.5905662 = 0.38943398 \) and \( |e_1| = 0.38943398 \).

- We can similarly work out the \( |e_i| \)'s for the remaining 119 observations in the sample, and verify that \( \sum_{i=1}^{120} e_i = 28.557784 \), as shown in the EXCEL output.
Multiple Linear Regression

- Now, let us test the significance of each explanatory variable.
Now, let us test the significance of each explanatory variable.

For example, for the variable IQ,

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \]

The test statistic, as shown in the output, is

\[ t = \frac{b_1 - 0}{S_{b_1}} = \frac{0.022195 - 0}{0.004904} = 4.5260 \]
Now, let us test the significance of each explanatory variable.

For example, for the variable IQ,

\[ H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \]

The test statistic, as shown in the output, is

\[ t = \frac{b_1 - 0}{S_{b_1}} = \frac{0.022195 - 0}{0.004904} = 4.5260 \]

The p-value, as reported in the output, is \( 1.49509 \times 10^{-5} \). Our test statistic clearly lies in the rejection region as it is smaller than \( \alpha = 0.05 \). Hence we reject \( H_0 \) and conclude that IQ is a significant variable.
Looking across the p-values of the t-tests for the six explanatory variables, we find that the tests for IQ, NoCrs, TravelMin, StudyHr and WorkHr all have p-values that do not exceed 0.05 and hence these variables are significant.
Multiple Linear Regression

- Looking across the p-values of the t-tests for the six explanatory variables, we find that the tests for IQ, NoCrs, TravelMin, StudyHr and WorkHr all have p-values that do not exceed 0.05 and hence these variables are significant.

- On the other hand, the t-tests for TravelMin and SleepHr have p-values of 0.289167731 and 0.614958993 respectively. These variables are therefore insignificant, with SleepHr having a higher p-value being the more insignificant of the two.
Multiple Linear Regression

- Looking across the p-values of the t-tests for the six explanatory variables, we find that the tests for IQ, NoCrs, TravelMin, StudyHr and WorkHr all have p-values that do not exceed 0.05 and hence these variables are significant.

- On the other hand, the t-tests for TravelMin and SleepHr have p-values of 0.289167731 and 0.614958993 respectively. These variables are therefore insignificant, with SleepHr having a higher p-value being the more insignificant of the two.

- We then remove TravelMin and SleepHr from the model. The new (reduced) model is

\[
SGPA = \beta_0 + \beta_1 IQ + \beta_2 NoCrs + \beta_5 StudyHr + \beta_6 WorkHr + u.
\]
Looking across the p-values of the t-tests for the six explanatory variables, we find that the tests for IQ, NoCrs, TravelMin, StudyHr and WorkHr all have p-values that do not exceed 0.05 and hence these variables are significant.

On the other hand, the t-tests for TravelMin and SleepHr have p-values of 0.289167731 and 0.614958993 respectively. These variables are therefore insignificant, with SleepHr having a higher p-value being the more insignificant of the two.

We then remove TravelMin and SleepHr from the model. The new (reduced) model is

\[ SGPA = \beta_0 + \beta_1 IQ + \beta_2 NoCrs + \beta_5 StudyHr + \beta_6 WorkHr + u. \]

The results, obtained using EXCEL, are as follows.
Multiple Linear Regression

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.145038169</td>
<td>0.589973932</td>
<td>1.94082841</td>
<td>0.05472641</td>
<td>-0.023586645</td>
<td>2.313662984</td>
</tr>
<tr>
<td>IQ</td>
<td>0.022513062</td>
<td>0.004863368</td>
<td>4.629109561</td>
<td>9.71844E-06</td>
<td>0.012879666</td>
<td>0.032146457</td>
</tr>
<tr>
<td>NoCrs</td>
<td>-0.055937296</td>
<td>0.02135401</td>
<td>-2.619521881</td>
<td>0.00999271</td>
<td>-0.098235479</td>
<td>-0.013639112</td>
</tr>
<tr>
<td>StudyHr</td>
<td>0.031518035</td>
<td>0.003999754</td>
<td>7.879992844</td>
<td>2.04156E-12</td>
<td>0.023595291</td>
<td>0.039440778</td>
</tr>
<tr>
<td>WorkHr</td>
<td>-0.030595661</td>
<td>0.00370361</td>
<td>-8.261038145</td>
<td>2.78543E-13</td>
<td>-0.037931799</td>
<td>-0.023259523</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

- All explanatory variables are significant.
Multiple Linear Regression

- All explanatory variables are significant.
- The coefficient estimates of the explanatory variables that remain in the model have changed (e.g., under the reduced model, $b_1 = 0.02251$ but under the previous full model, $b_1 = 0.02219$). Note that Least squares estimates take into account the multiple correlation among the explanatory variables, which changes as variables are added or dropped.
Multiple Linear Regression

- All explanatory variables are significant.
- The coefficient estimates of the explanatory variables that remain in the model have changed (e.g., under the reduced model, $b_1 = 0.02251$ but under the previous full model, $b_1 = 0.02219$). Note that Least squares estimates take into account the multiple correlation among the explanatory variables, which changes as variables are added or dropped.
- $R^2$ has dropped from 0.90332 to 0.90195. Note that $R^2$ always decreases as variables are removed but this does not always mean the reduced model is inferior to the full model.
Multiple Linear Regression

- All explanatory variables are significant.
- The coefficient estimates of the explanatory variables that remain in the model have changed (e.g., under the reduced model, $b_1 = 0.02251$ but under the previous full model, $b_1 = 0.02219$). Note that Least squares estimates take into account the multiple correlation among the explanatory variables, which changes as variables are added or dropped.
- $R^2$ has dropped from 0.90332 to 0.90195. Note that $R^2$ always decreases as variables are removed but this does not always mean the reduced model is inferior to the full model.
- It can be verified that $\sum_{i=1}^{120} |e_i| = 28.44398759$, as shown in the EXCEL output. This sum of absolute errors is slightly smaller than that under the full model.