Scatter Plot, Correlation

&

Regression Line

What is a **Scatter Plot**?

A **Scatter Plot** is a plot of ordered–pairs (x, y) where the horizontal axis is used for the x variable and the vertical axis is used for the y variable.

How is **Scatter Plot** helpful?

The pattern of the plotted points in a **Scatter Plot** will enable us to see whether there is a relationship between the two variables.

Example:

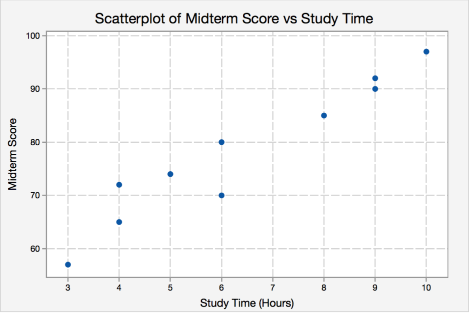
The study time and midterm exam score for a random sample of 10 students in a statistic course are shown in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Draw the scatter plot.

Solution:

We plot the ordered-pair $(3, 57)$ for student A, ordered-pair $(4, 65)$ for student B, and so on to draw the **Scatter Plot**.



What is the **Regression Line**?

The **Regression Line** is the graph of the **Regression Equation**.

What is a **Regression Equation**?

The **Regression Equation** algebraically describes the best linear relationship between two variables x and y. The **Regression Equation** is usually written in the following form.

$$\hat{y} = a + bx$$

How do we compute a and b?

- Compute $\sum x$, $\sum y$, and $\sum xy$.
- Compute $\sum x^2$, and $\sum y^2$.
- Now we use the formulas below.

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

and

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

where n is the number of ordered-pairs.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistic course are shown in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Find the equation of the regression line in which the x variable is the study time, and y variable is the midterm score. Draw the regression line and scatter plot in the same coordinate system.

Solution:

We first identify that $n = 10$, then find and verify that $\sum x = 64$, $\sum y = 782$, $\sum xy = 5277$, $\sum x^2 = 464$, $\sum y^2 = 62632$,

and then we apply these values in the the formula

$$\begin{aligned} b &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \\ &= \frac{10(5277) - (64)(782)}{10(464) - (64)^2} \\ &= \frac{2722}{544} \\ &\approx 5.004 \end{aligned}$$

Solution Continued:

and

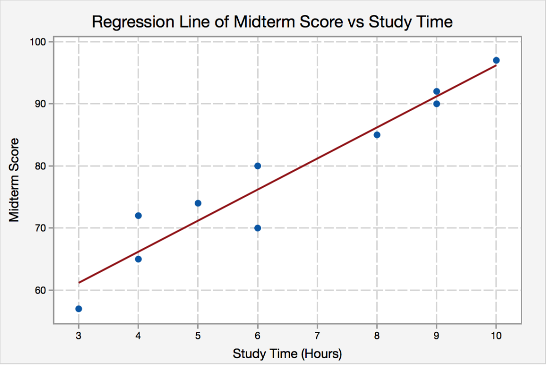
$$
\begin{aligned}
a &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \\
&= \frac{(782)(464) - (64)(5277)}{10(464) - (64)^2} \\
&= \frac{25120}{544} \\
&\approx 46.176
\end{aligned}
$$

So the equation of the regression line is

$$
\begin{aligned}
\hat{y} &= a + bx \\
&= 46.176 + 5.004x
\end{aligned}
$$

Solution Continued:

Here is the graph of the regression line as well as the scatter plot.



What is a **Correlation**?

A **Correlation** between two variables is when there is an apparent association between the values of one variable with the corresponding values from the other variable.

What is the **Linear Correlation Coefficient**?

The **Linear Correlation Coefficient** is a numerical value that measures the strength of the linear correlation between the paired x and y for all values in the sample. We denote this value by r.

What are the properties of r?

- $-1 \leq r \leq 1$
- It is not designed to measure the strength of a nonlinear relationship.
- It is very sensitive and changes value if the sample contains any outliers.
- The **Linear Correlation Coefficient** is considered **significant** when $\left|r\right|$ is fairly close to 1.

How do we compute r?

- Compute $\sum x$, $\sum y$, and $\sum xy$.
- Compute $\sum x^2$, and $\sum y^2$.
- Now we use the formula

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

where n is the number of ordered-pairs.

It is worth noting that r is usually calculated with a computer software or a calculator.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Find the value of linear correlation coefficient r.

Solution:

We first identify that $n = 10$, then find and verify that
$\sum x = 64$, $\sum y = 782$, $\sum xy = 5277$, $\sum x^2 = 464$,
$\sum y^2 = 62632$,
and then we apply these values in the the formula

$$\begin{aligned}
r &= \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \\
&= \frac{10(5277) - (64)(782)}{\sqrt{10(464) - (64)^2}\sqrt{10(62632) - (782)^2}} \\
&= \frac{2722}{\sqrt{544}\sqrt{14796}} \\
&\approx 0.959
\end{aligned}$$

What is the **Coefficient of Determination**?

The **Coefficient of Determination** is a numerical value usually provided in percentage that indicates what percentage of the dependent variable y is explained by the independent variable x. We denote this value by r^2.

How do we compute r^2?

We simply square the value of r and then convert it to a percentage by moving the decimal point two places to the right.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Find the value of coefficient of determination r^2 and explain what this number describes in the context of this example.

Solution:

We have already used this example and found the value of the linear correlation coefficient r.

We got $r \approx 0.959$.

Now we square this number to get the coefficient of determination.

$$\begin{aligned} r^2 &= (0.959)^2 \\ &= 0.919681 \\ &\approx 0.920 \\ &\approx 92.0\% \end{aligned}$$

So 92.0% of the midterm scores are explained by the study time.

Example:

The study time and midterm exam score for a random sample of 10 students in a statistics course are shown in the following table.

Student	A	B	C	D	E	F	G	H	I	J
Study Time; x (Hours)	3	4	4	5	6	6	8	9	10	9
Midterm Score; y	57	65	72	74	70	80	85	90	97	92

Find the value of the expression $r \cdot \sqrt{\dfrac{n-2}{1-r^2}}$, rounded to 3–decimal places when needed.

Solution:

We already know that $n = 10$, and have computed
$r = 0.959$, and $r^2 = 0.920$,

now we apply these values to the expression we need to compute.

$$
\begin{aligned}
r \cdot \sqrt{\frac{n-2}{1-r^2}} &= 0.959 \cdot \sqrt{\frac{10-2}{1-0.920}} \\
&= 0.959 \cdot \sqrt{\frac{8}{0.08}} \\
&= 0.959 \cdot \sqrt{100} \\
&= 0.959 \cdot 10 \\
&= 9.59
\end{aligned}
$$

How do we make **prediction**?

- When **linear correlation is significant,** use $\hat{y} = a + bx$.

  Plug in the given x value to find the prediction value y.

- When **linear correlation is not significant,** use $\overline{y}$.

Example:

Eight pairs of data yield the regression line equation
$\hat{y} = 55.6 + 2.8x$ with $\bar{y} = 71.5$.
What is the best predicted value for y for $x = 5.5$ if we assume the linear correlation is significant?

Solution:

Since the linear correlation coefficient is significant, we use the equation of the regression line $\hat{y} = 55.6 + 2.8x$. and plug in $x = 5.5$ to find the prediction value.

$$\begin{aligned} \hat{y} &= 55.6 + 2.8x \\ &= 55.6 + 2.8(5.5) \\ &= 55.6 + 15.4 \\ &= 71 \end{aligned}$$

So, our prediction value is 71.

Example:

Ten pairs of data yield the regression line equation
$\hat{y} = 73.5 - 4.5x$ with $\bar{y} = 58.5$.

What is the best predicted value for y for $x = 4.5$ if we assume the linear correlation is not significant?

Solution:

Since the linear correlation coefficient is not significant,

we use $\bar{y}$ as the prediction value regardless of the value of x.

So, our prediction value is 58.5.

Causation vs Correlation