Correlation

1. Are two or more variables related?

2. If yes, what is the strength of the relationship?

3. What type of relationship exists?

4. What kind of predictions can be made from the relationship?

Correlation is a statistical method used to determine whether a linear relationship between variables exists.

Correlation is defined as a measure of the linear relationship between two quantitative variables.

A Correlation is a single number that describes the degree of relationship between two variables.

1. Are two or more variables related?
2. If so, what is the strength of the relationship?

To answer these two questions, statisticians use the correlation coefficient, a numerical measure to determine whether two or more variables are related and to determine the strength of the relationship between or among the variables.

3. What type of relationship exists?

There are two types of relationships: simple and multiple.

In a simple relationship, there are two variables: an independent variable (explanatory variable or predictor variable) and a dependent variable (response variable).

In a multiple relationship, there are two or more independent variables that are used to predict one dependent variable.

4. What kind of predictions can be made from the relationship?

We make predictions in our daily life. Examples include weather forecasting, stock market analyses, sales predictions, crop predictions, gasoline price predictions, and sports predictions.

Scatter Plots and Correlation

- The independent and dependent variables can be plotted on a graph called a scatter plot.

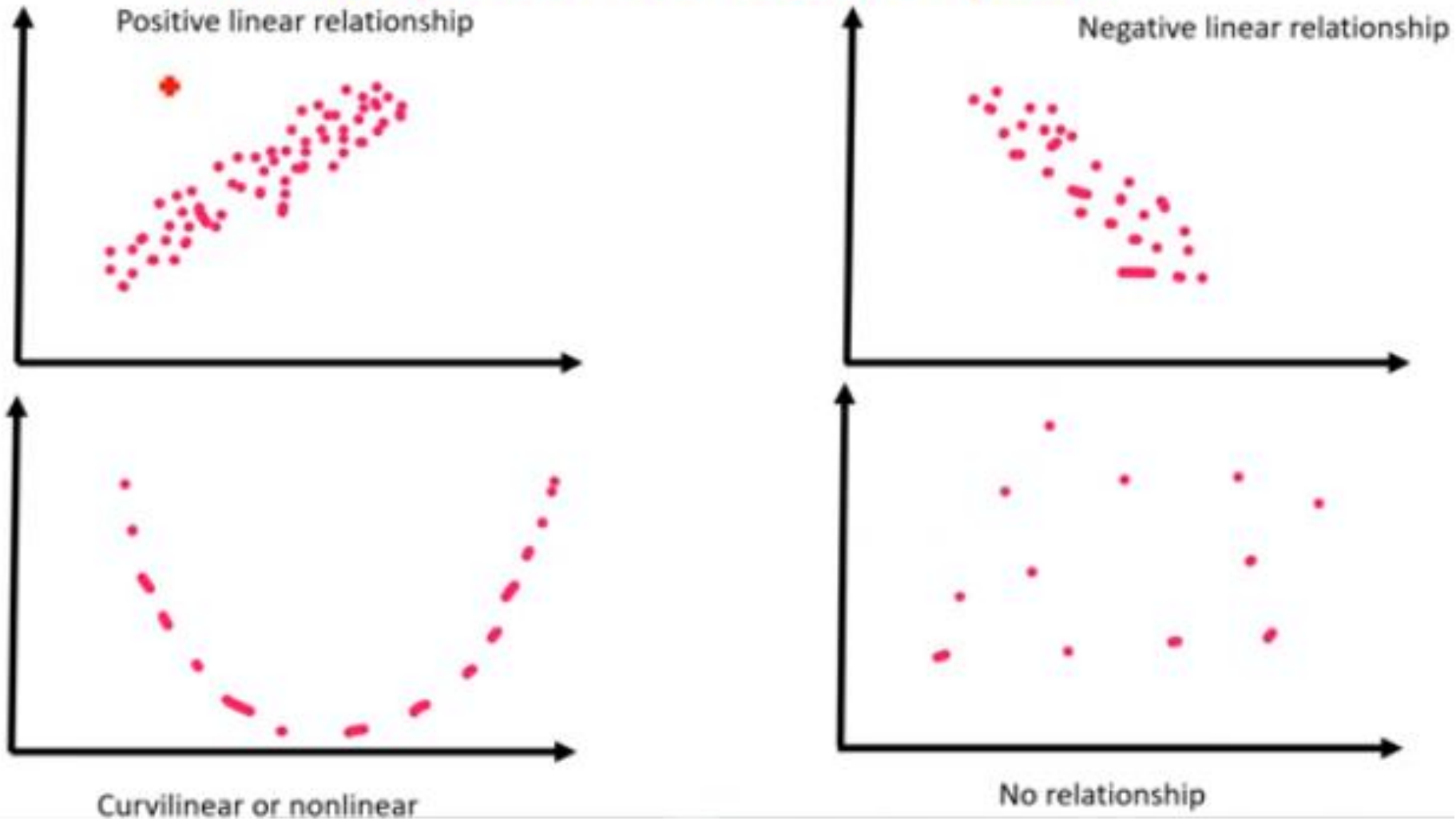
- A scatterplot is a type of data display that shows the relationship between two numerical variables.(independent x axis, and dependent variable y axis.)

- Independent variable can be controlled while depended variable can not be controlled.

e.g. attendance and grade

control no control

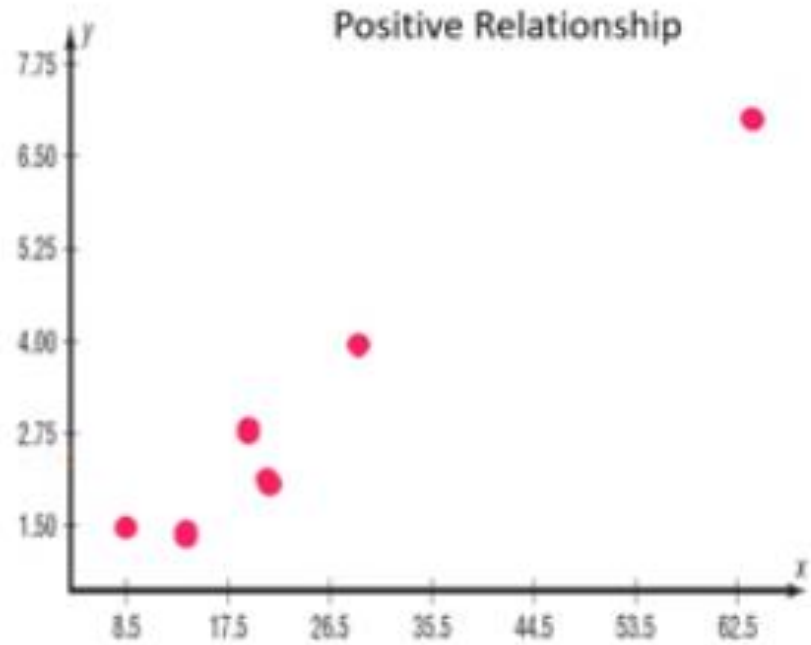
Various patterns of scatter plots



Car Rental Companies

Construct a scatter plot for the following data.

Company	Cars (in ten thousands)	Revenue(in Billions)
A	20.8	2.1
B	63	7
C	29	3.9
D	8.5	1.5
E	13.4	1.4
F	19.1	2.8

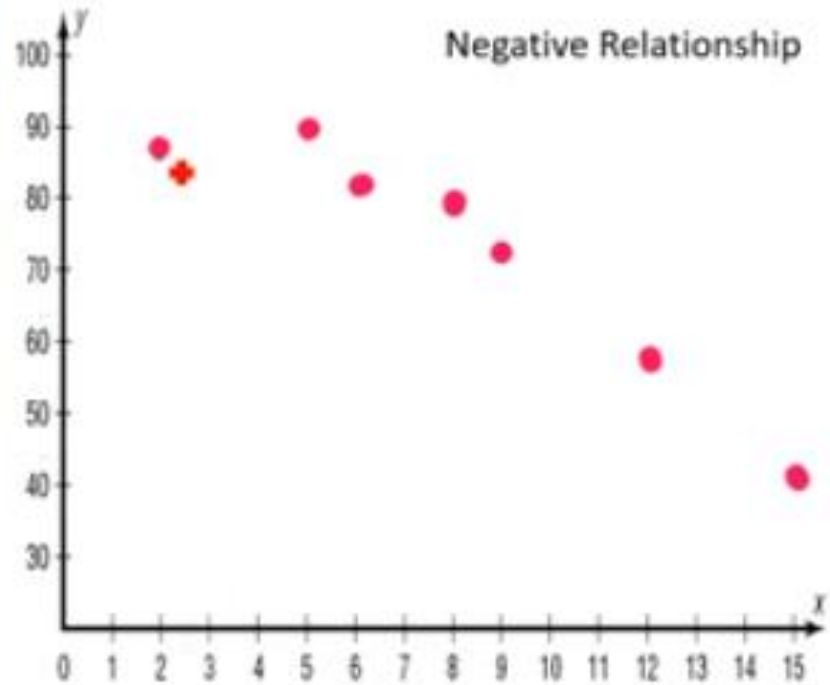


Draw and label the *x* and *y* axes.
Plot each point on the graph.
Determine the type of relationship that exits for the variables.

Absences/Final Grades

Construct a scatter plot for the number of absences and the final grades of seven students from a Math class.

Student	Number of absences x	Final grade y(%)
T	2	86
U	6	82
V	15	43
W	12	58
X	9	74
Y	5	90
Z	8	78



Draw and label the *x* and *y* axes.

Plot each point on the graph.

Determine the type of relationship that exits for the variables.

Age and Money

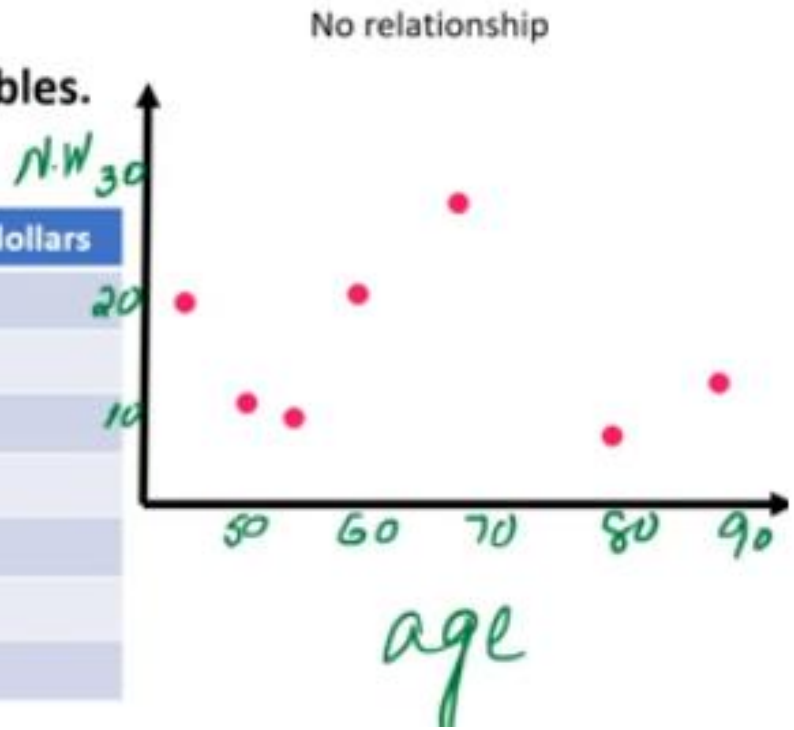
To find if there is a relationship between the ages of the people and their net worth.

Draw and label the *x* and *y* axes.

Plot each point on the graph.

Determine the type of relationship that exits for the variables.

Person	Age X	Net Worth Y in billions of dollars
A	50	12
B	67	21
C	90	13
D	45	20
E	70	25
F	80	7
G	55	10



Correlation Coefficient are used to determine the strength of the linear relationship between two variables.

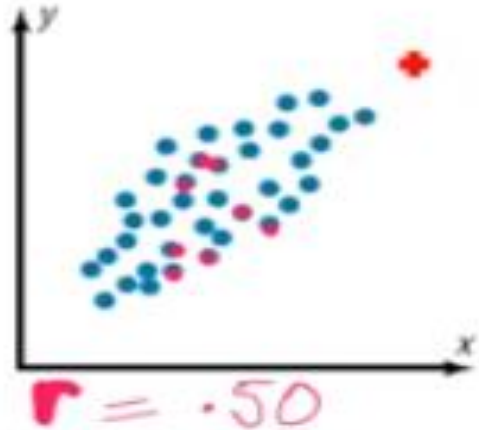
- The population correlation coefficient, denoted by ρ, is computed by using all possible pairs of data values (x, y) taken from a population.

- The linear correlation coefficient, denoted by r, is computed from the sample data and measures the strength and direction of a linear relationship between two quantitative variables. Which is called Pearson product moment correlation coefficient (PPMC).

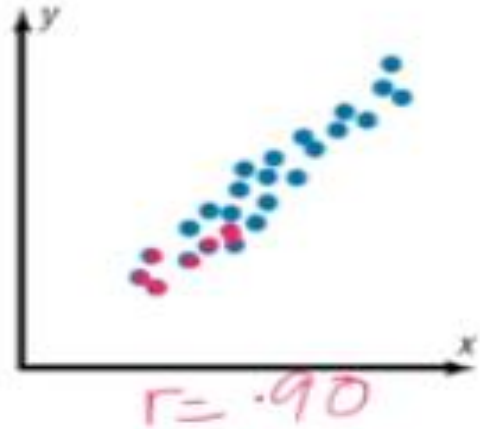
- The range of the correlation coefficient is from –1 to +1.

linear correlation coefficient

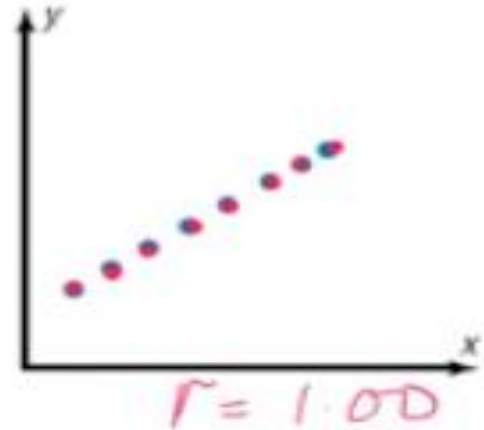
-1 to +1



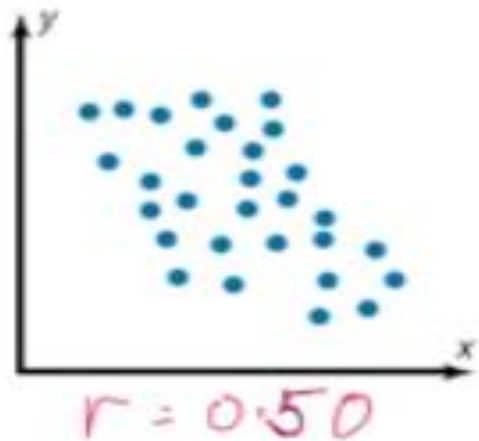
$r = .50$



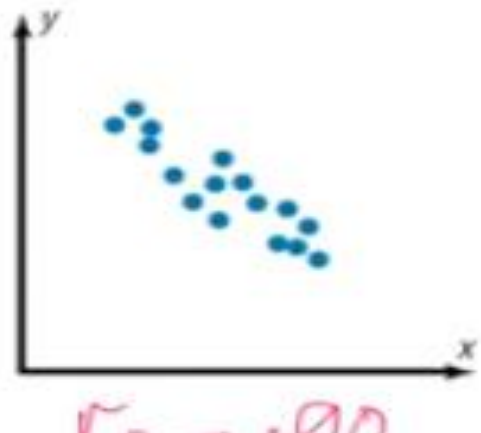
$r = .90$



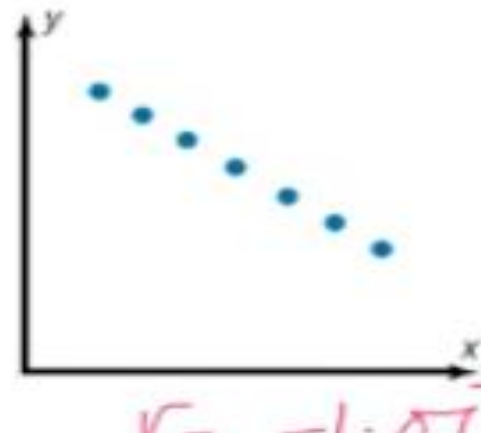
$r = 1.00$



$r = 0.50$



$r = -.90$



$r = -1.00$

formula for the correlation coefficient r

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

where *n* is the number of data pairs.

Compute the correlation coefficient r for the following data

63x63

Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
U	63.0	7.0	441.00	3969.00	49.00
V	29.0	3.9	113.10	841.00	15.21
W	20.8	2.1	43.68	432.64	4.41
X	19.1	2.8	53.48	364.81	7.84
Y	13.4	1.4	18.76	179.56	1.96
Z	8.5	1.5	12.75	72.25	2.25
$n=6$	$\Sigma x =$ 153.8	$\Sigma y =$ 18.7	$\Sigma xy =$ 682.77	$\Sigma x^2 =$ 5859.26	$\Sigma y^2 =$ 80.67

$\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, $\Sigma x^2 = 5859.26$, $\Sigma y^2 = 80.67$, $n = 6$

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

$$r = \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{\left[(6)(5859.26) - (153.8)^2\right]\left[(6)(80.67) - (18.7)^2\right]}}$$

$r = 0.982$ (strong positive relationship)

Practice Problem:

Student	Number of absences x	Final grade y(%)
T	2	86
U	6	82
V	15	43
W	12	58
X	9	74
Y	5	90
Z	8	78

$n = 7$

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

$$r = \frac{(7)(3745) - (57)(511)}{\sqrt{\left[(7)(579) - (57)^2\right]\left[(7)(38{,}993) - (511)^2\right]}}$$

$r = -0.944$ (strong negative relationship)

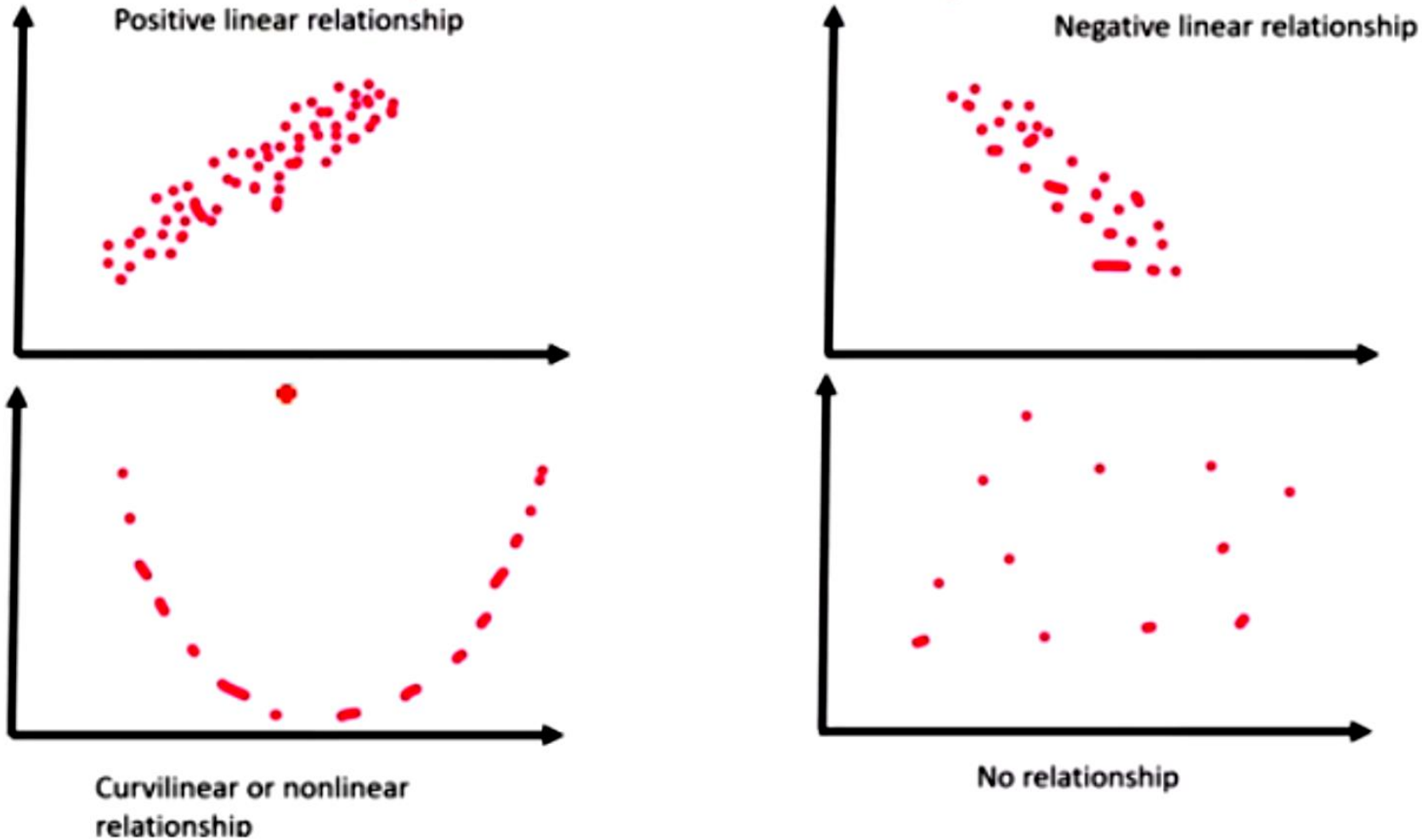
Practice Problem

Model	Price	Engine Size
Toyota	3000	50
Honda	10000	80
Toyota	15000	100
Mazda	20000	120
Toyota	23000	140
Honda	25000	160
Mazda	30000	180
Audi	26000	150
BMW	45000	250
Honda	28000	160
Audi	24000	190
BMW	55000	300
Honda	28000	160

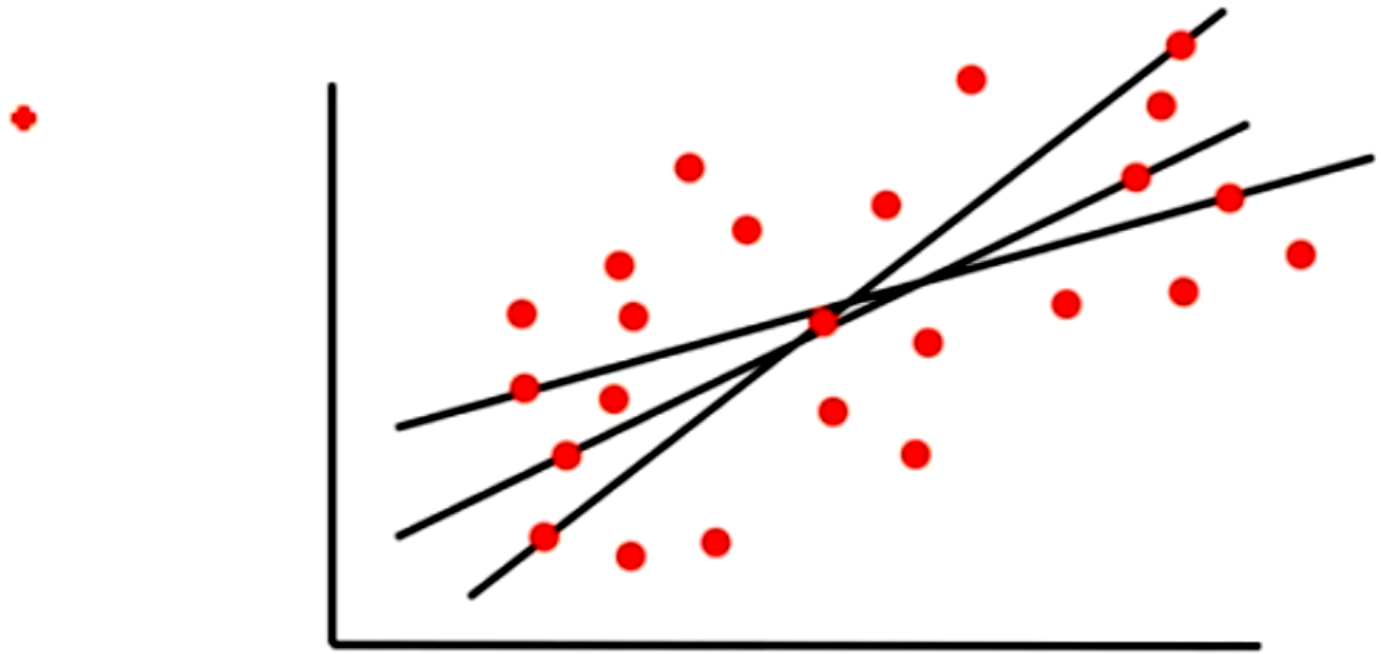
0.98 +ve correlation

Regression

Various patterns of scatter plots

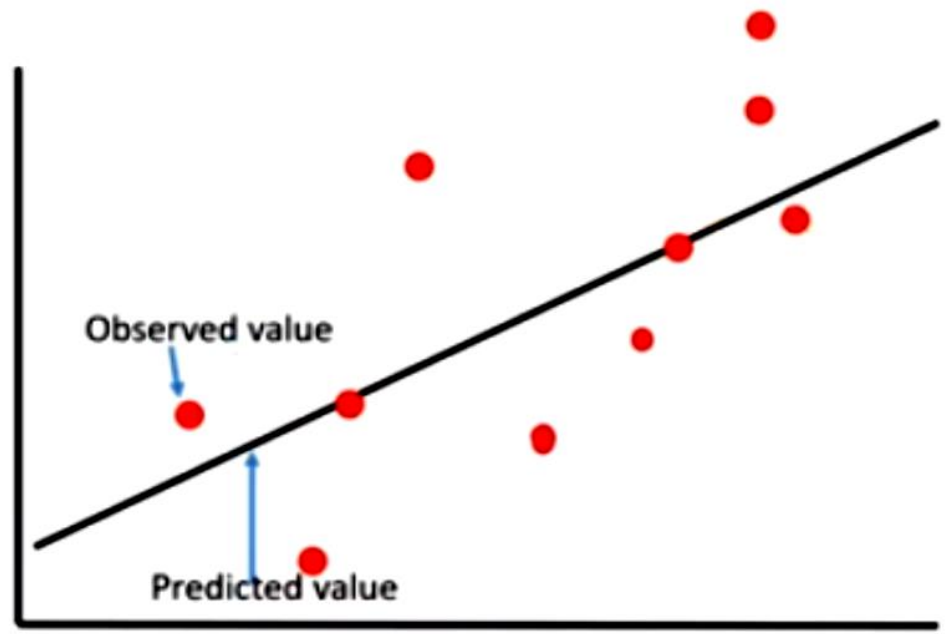


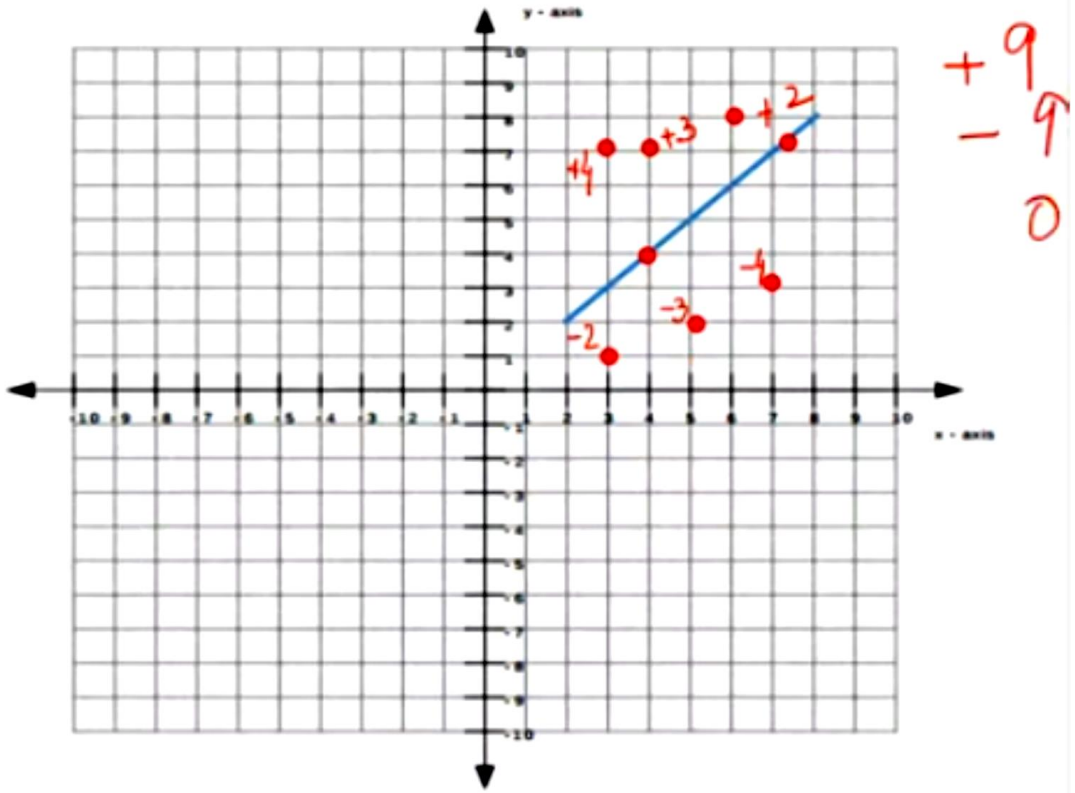
- If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line** which is the data's line of best fit.

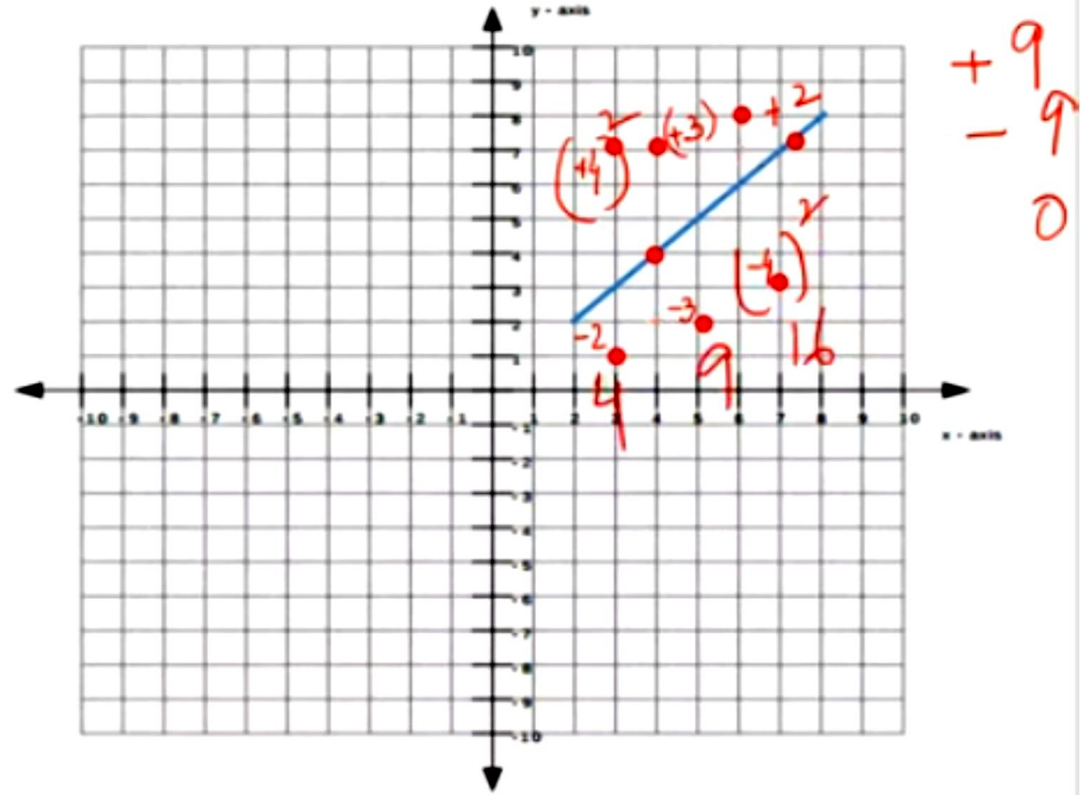


Note: Determining the regression line when (r) correlation coefficient is not significant and making predictions using the regression line are pointless.

Line of best fit: Best fit means that the sum of the squares of the vertical distances from each point to the line is at minimum.




y - axis
x - axis
+4
+3
+2
-2
-3
-4
+9
−9
0


y - axis
x - axis
(+4)²
(+3)²
+2
(-2)²
(-3)²
(-4)²
16
9
4
+9
−9
0

Regression equation

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

Total product = fixed + cost per unit(number of units)

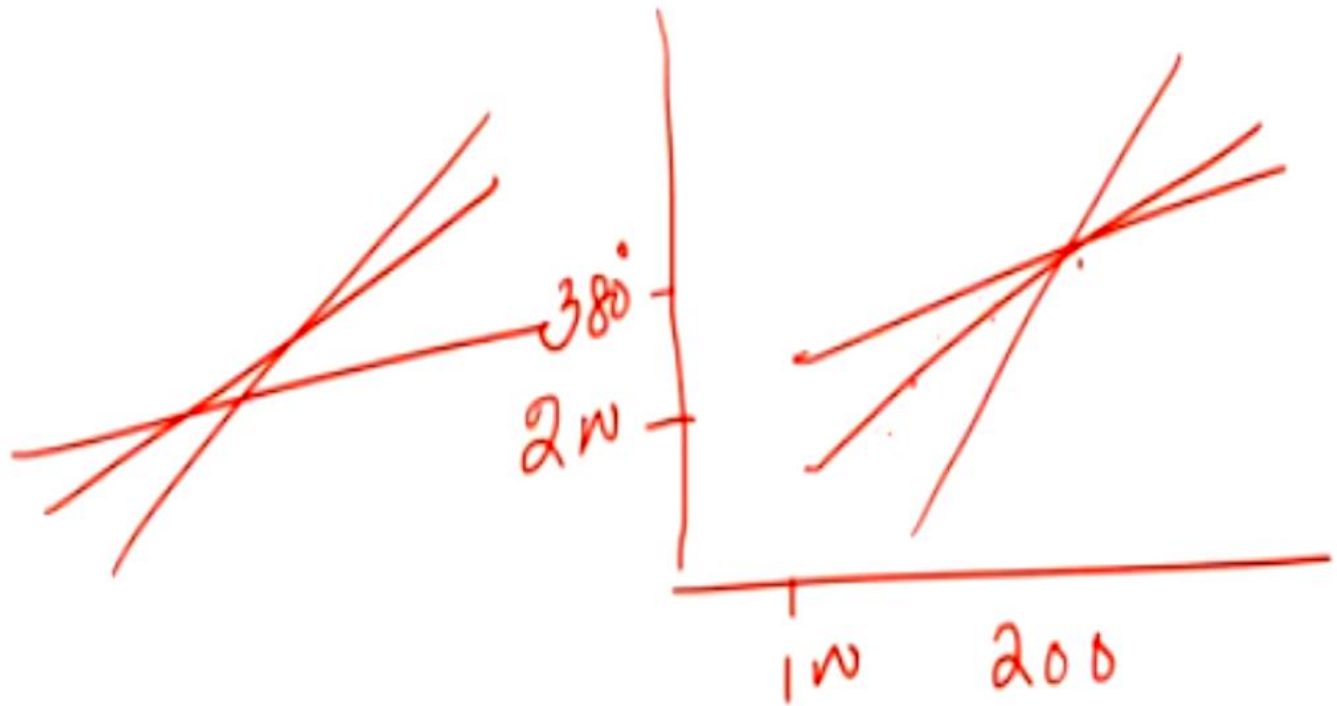
The difference between the actual value of y and the predicted value y is called a residual or predicted error.

Residuals are used to determine the line that best describes the relationship between two variables.

The method used for making residuals as small as possible is called the method of least squares.

The regression line is also called the **least square regression line.**

The reason you need a line of best fit is that the values of y will be predicted from the values of x, the closer the points are to the line, the better the fit and the prediction will be.



Regression Line

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

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

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

$a = \hat{y}$ intercept

b = slope of the line

$\frac{\text{rise}}{\text{run}}$

$$b = \sum xy / \sum x^2$$

$$a = \overline{Y} - b\overline{X}$$

• Finding the Regression Line Equation

1. **Make a table, as shown below.**
2. **Find the values of xy, x^2, and y^2 and sum each column.**

Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
U	63.0	7.0			
V	29.0	3.9			
W	20.8	2.1			
X	19.1	2.8			
Y	13.4	1.4			
Z	8.5	1.5			

$\hat{y} = a + b x$

Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
U	63.0	7.0	441.00	3969.00	49.00
V	29.0	3.9	113.10	841.00	15.21
W	20.8	2.1	43.68	432.64	4.41
X	19.1	2.8	53.48	364.81	7.84
Y	13.4	1.4	18.76	179.56	1.96
Z	8.5	1.5	12.75	72.25	2.25
	$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$
$n=6$	153.8	18.7	682.77	5859.26	80.67

Finding the Correlation Coefficient and the Regression Line Equation

3. Substitute in the formula to find the value of r.

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

4 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $\hat{y} = a + bx$.

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

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

Car Rental Companies

Find the equation of the regression line and graph the line on the scatter plot.

$\Sigma x = 153.8, \quad \Sigma y = 18.7, \quad \Sigma xy = 682.77, \quad \Sigma x^2 = 5859.26, \quad \Sigma y^2 = 80.67, \; n = 6$

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

$$= \frac{(18.7)(5859.26) - (153.8)(682.77)}{6(5859.26) - (153.8)^2} = 0.396$$

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

$$= \frac{6(682.77) - (153.8)(18.7)}{6(5859.26) - (153.8)^2} = 0.106$$

$\hat{y} = a + bx \rightarrow \hat{y} = 0.396 + 0.106x$

Find two points to sketch the graph of the regression line.

Use any *x* values. For example, let *x* equal 10 and 30. Substitute in the equation and find the corresponding *y* value.

Plot (10,1.456) and (30,3.576), and sketch the resulting line.

$\hat{y} = 0.396 + 0.106x$

$= 0.396 + 0.106(10)$

$= 0.396 + 1.06$

$\hat{y} = 1.456$

$(10, 1.456)$

$\hat{y} = 0.396 + .106x$

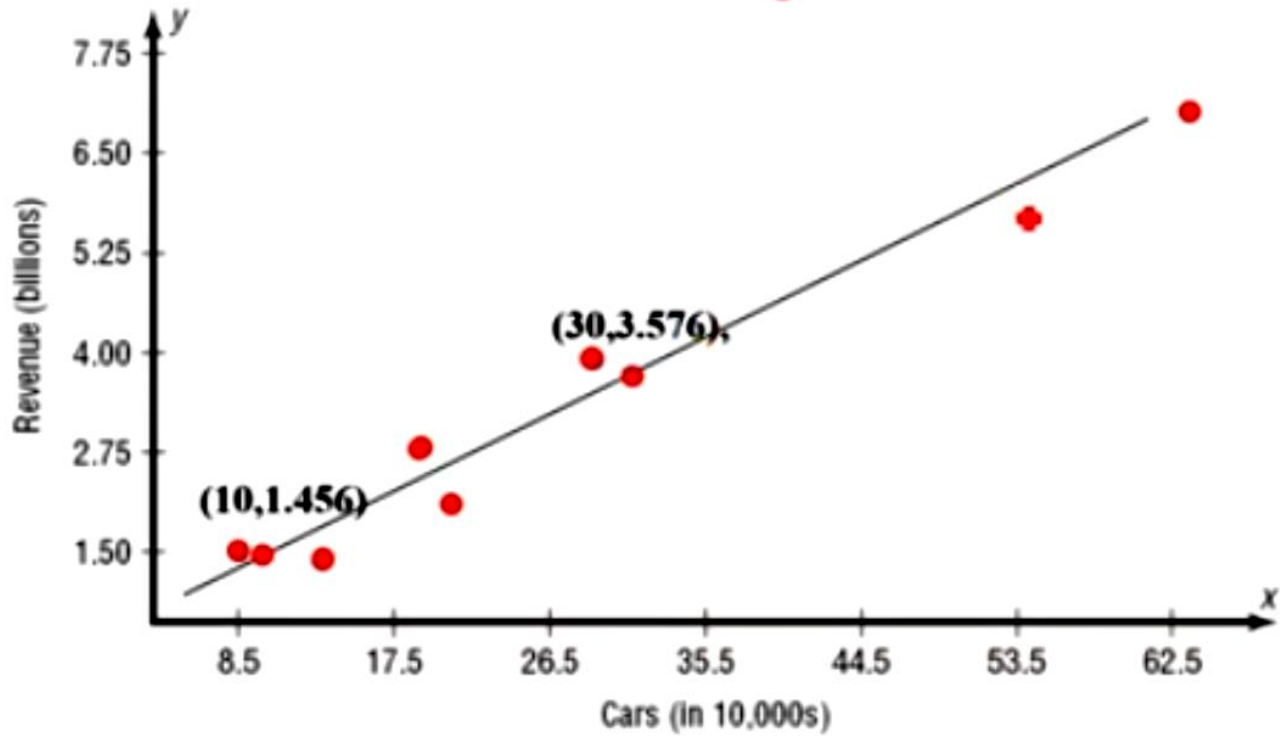
$= 0.396 + 0.106(30)$

$= 0.396 + 3.18$

$\hat{y} = 3.576$

Find the equation of the regression line and graph the line on the scatter plot.

$\hat{y} = 0.396 + 0.106x$



Car Rental Companies

Use the equation of the regression line to predict the income of a car rental agency that has 300,000 automobiles.

x = 30 corresponds to 300,000 automobiles.

$$\hat{y} = 0.396 + 0.106x$$
$$\hat{y} = 0.396 + 0.106(30)$$
$$= 0.396 + 3.18$$
$$= 3.576$$

When a rental agency has 300,000 automobiles, its revenue will be approx.: $3.576 billion.