7. Maths - Statistics – PART – 7

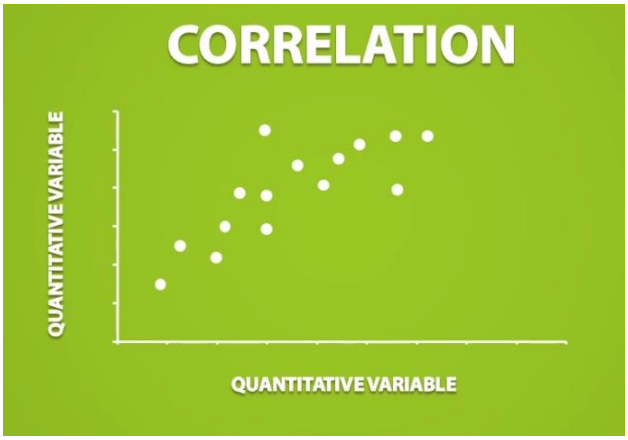
Contents

7. Maths - Statistics – PART – 7

✓ In this chapter we will discuss about Regression and R-Squared

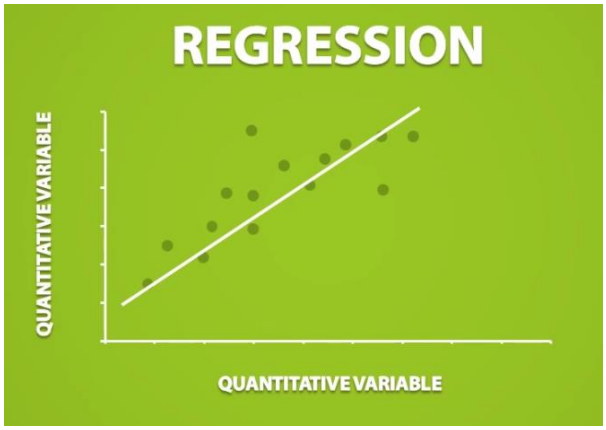
1. Correlation

✓ Correlation explains about how we measure the direction and strength of linear relationship in between quantitative variables



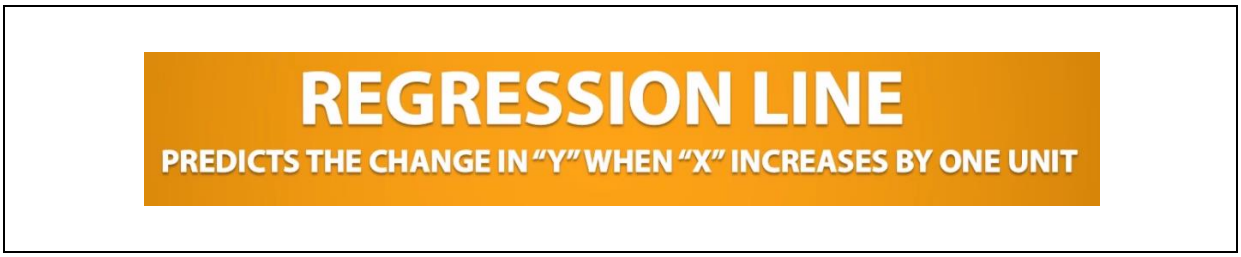
2. Regression

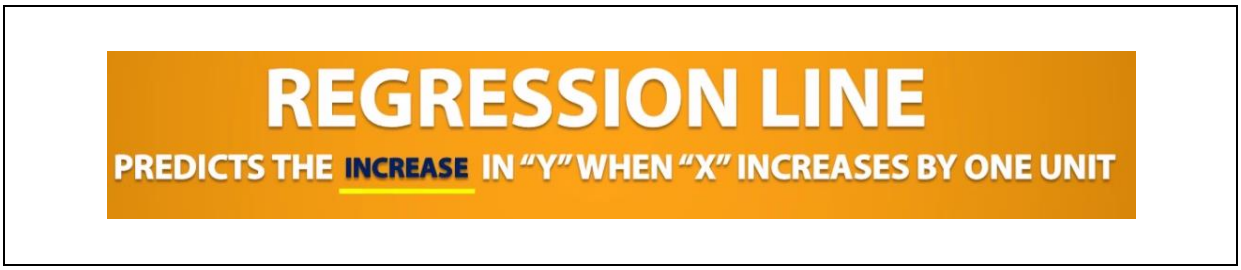
- ✓ Regression explains about how we can draw line in between the points
- ✓ This line represents about the pattern of the data
- ✓ This line is called as regression line

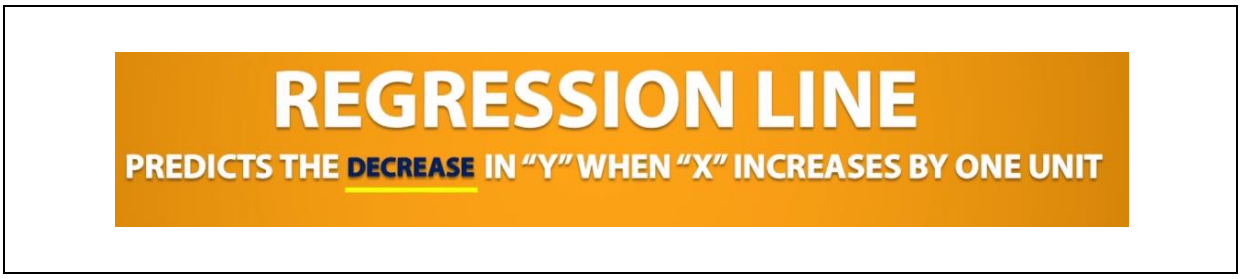


3. Regression line

- ✓ Regression line predicts the change in **Y** when **X** increases by one unit
- ✓ Here change in Y can be increase/decrease

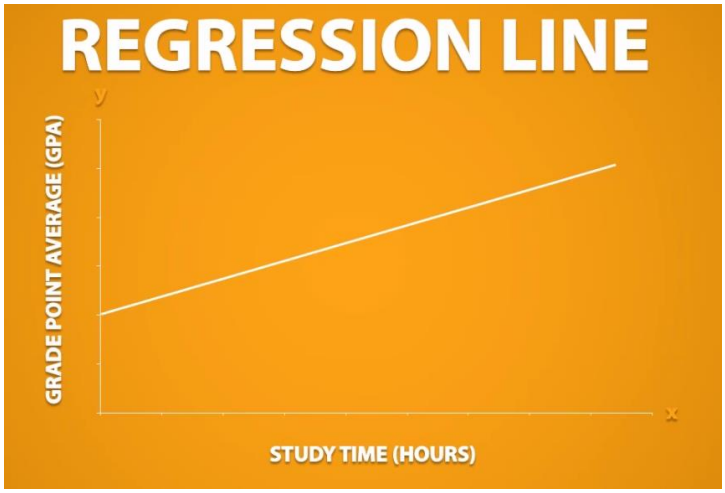




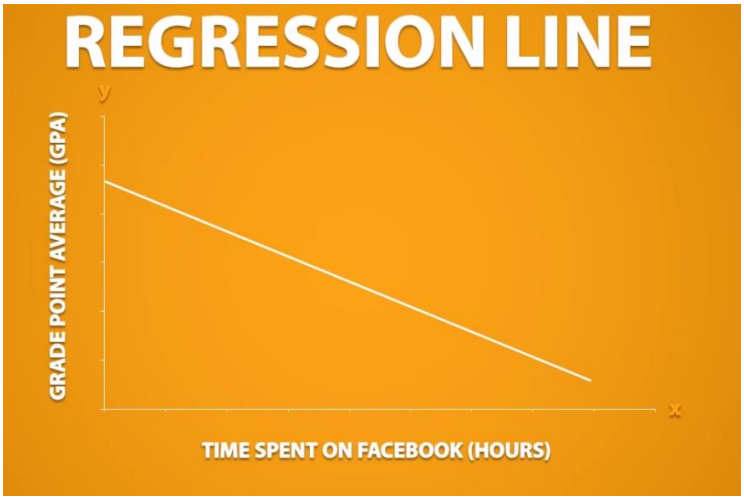


3.1. Example: Regression line

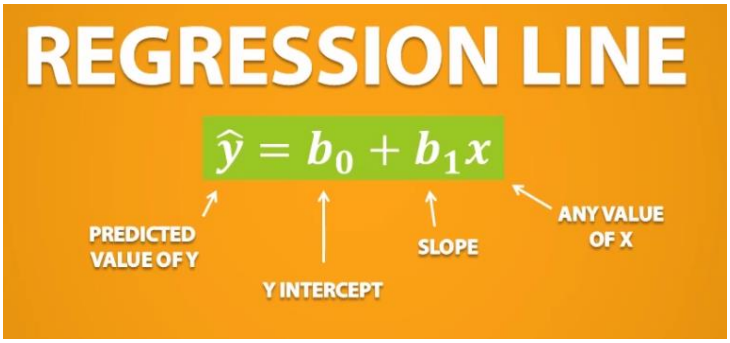
- ✓ If we study time (hours) on X axis and grade on Y axis then we should access some positive relationship in between these two variables.
- ✓ Generally speaking the more you study then we will get good grade ☺

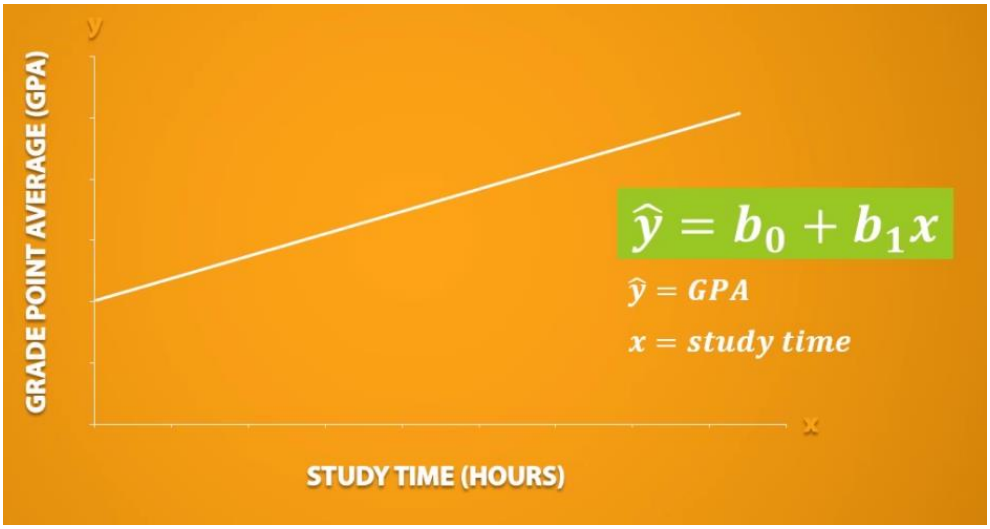


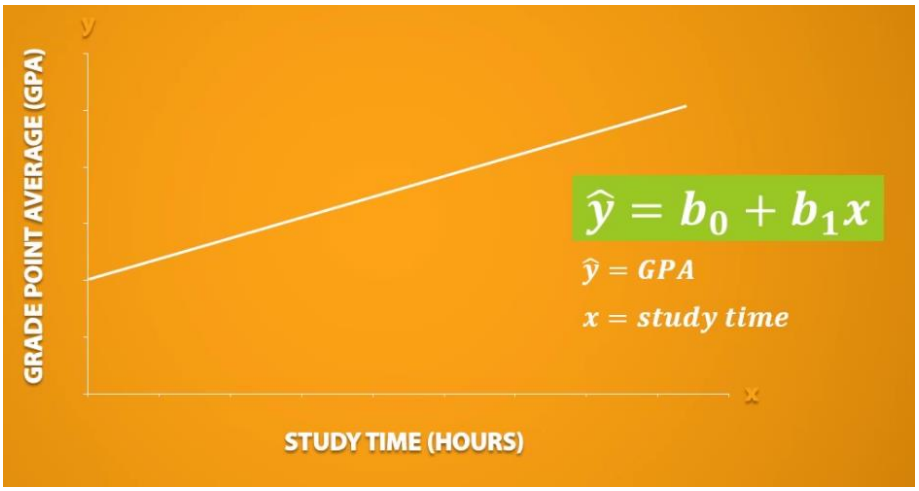
- ✓ Instead of studying if we spend more time on Facebook then we will get negative relationship

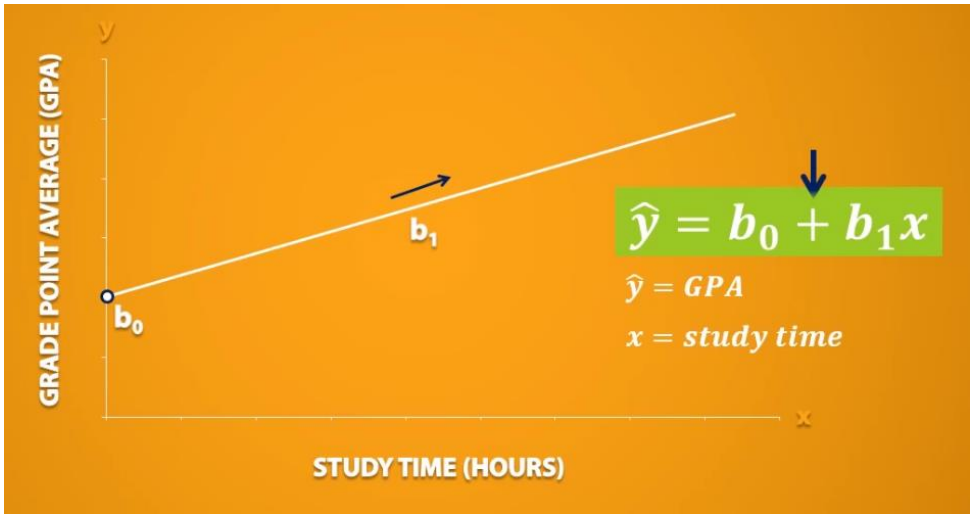


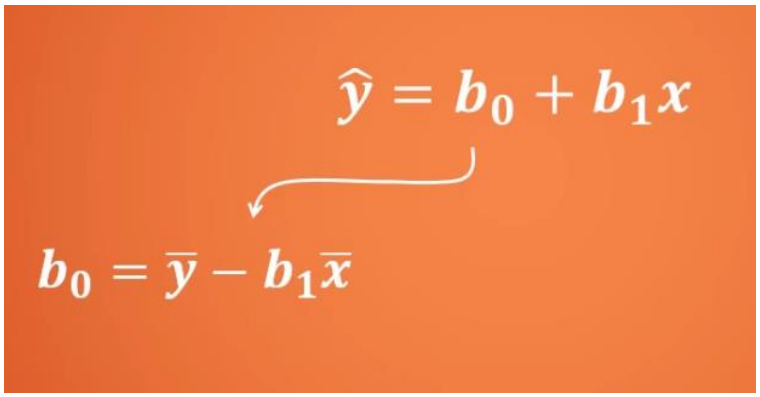
4.1. Regression line formula

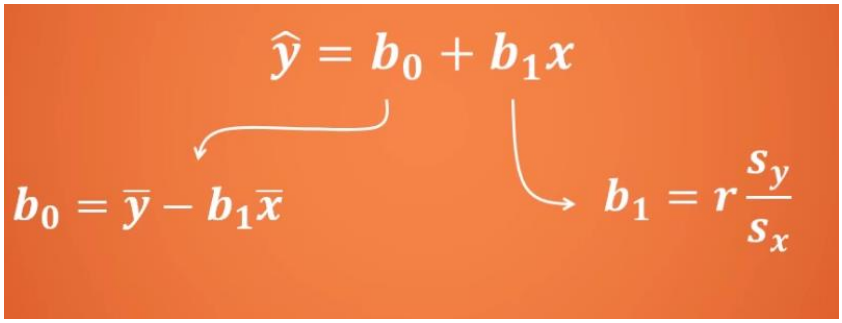





y
GRADE POINT AVERAGE (GPA)
$\hat{y} = b_0 + b_1 x$
$\hat{y} = GPA$
$x = study\ time$
x
STUDY TIME (HOURS)


y
GRADE POINT AVERAGE (GPA)
b_1
b_0
$\hat{y} = b_0 + b_1 x$
$\hat{y} = GPA$
$x = study\ time$
x
STUDY TIME (HOURS)

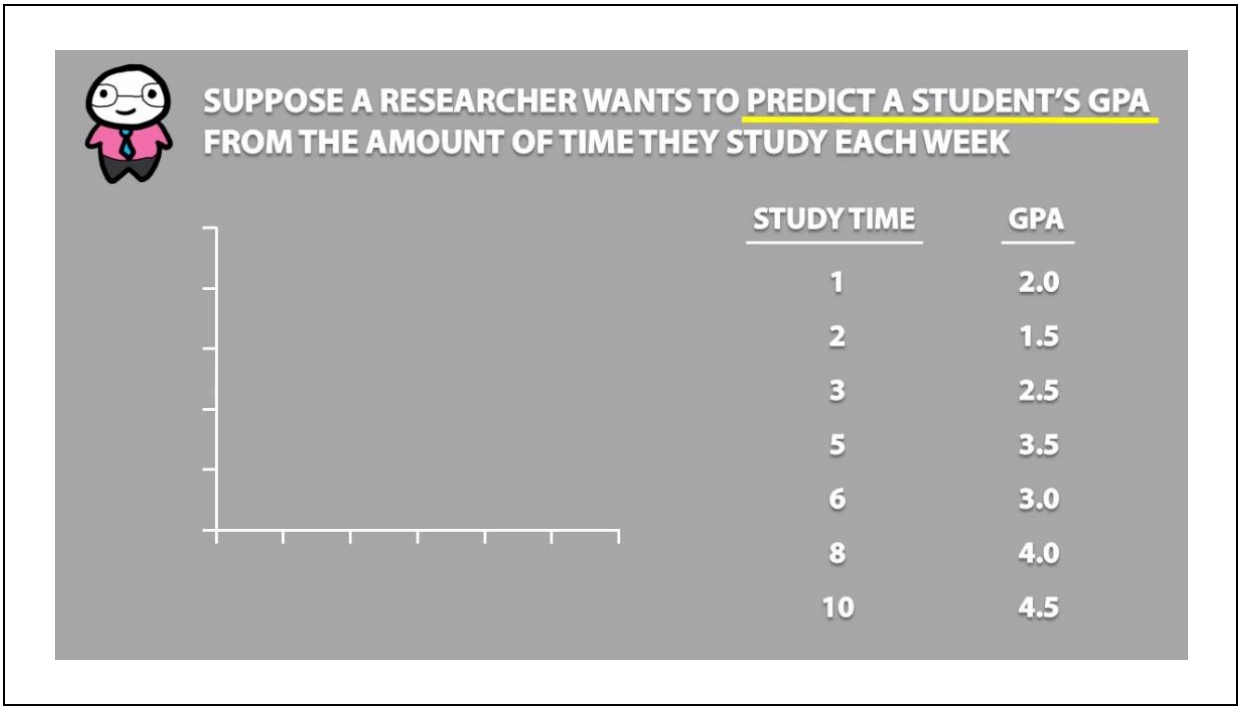

$\hat{y} = b_0 + b_1 x$
$b_0 = \bar{y} - b_1 \bar{x}$

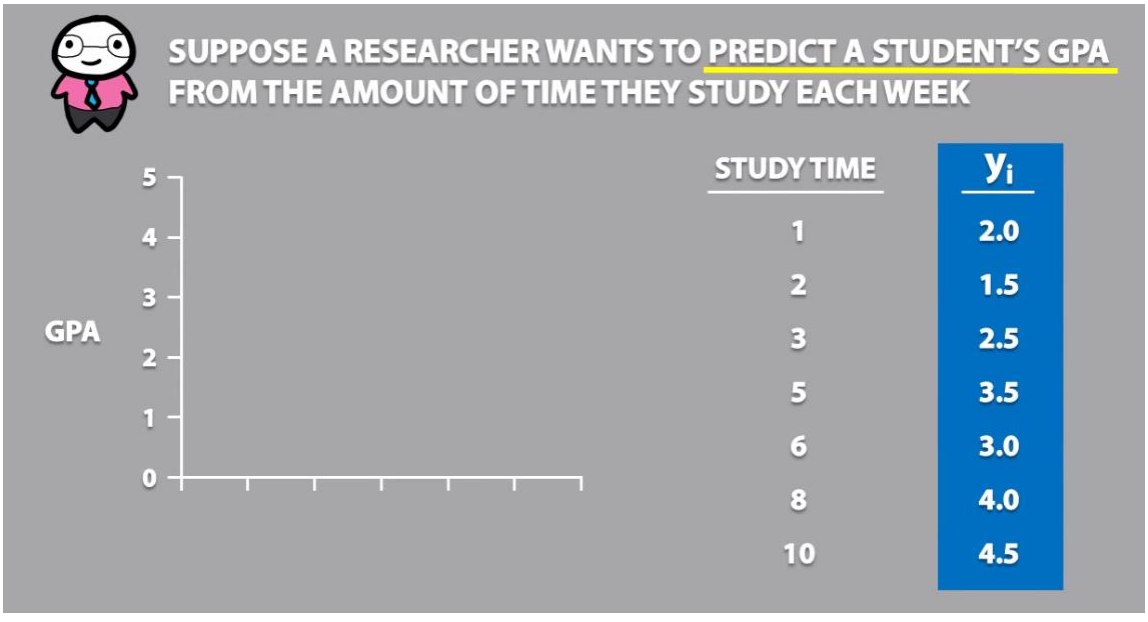

$\hat{y} = b_0 + b_1 x$
$b_0 = \bar{y} - b_1 \bar{x}$
$b_1 = r \frac{s_y}{s_x}$

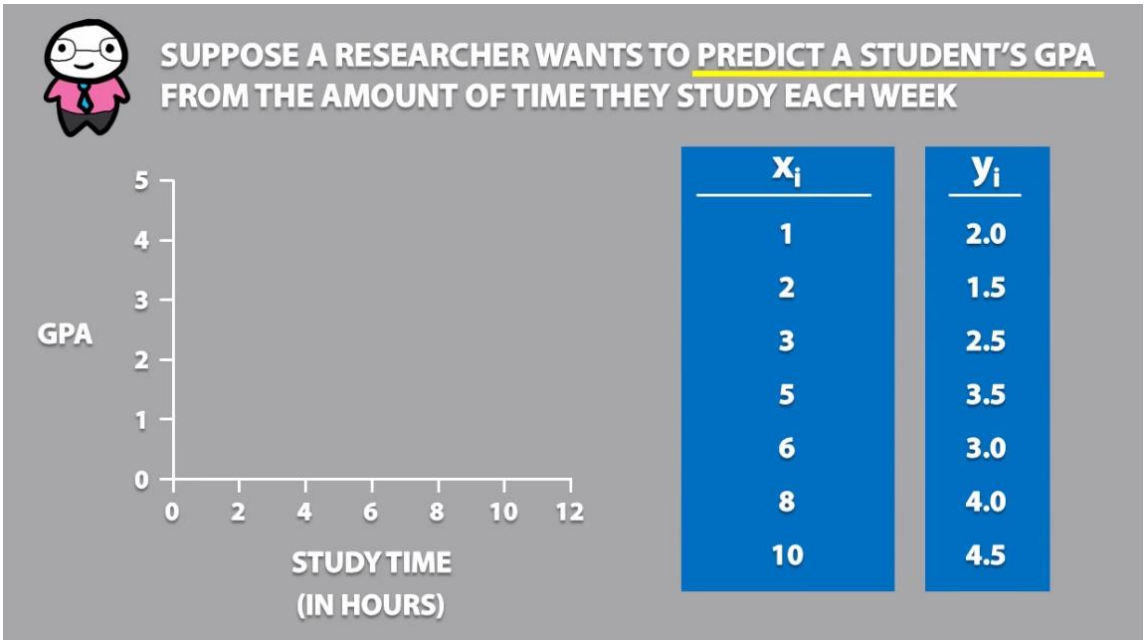
Scenario


SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA
FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK

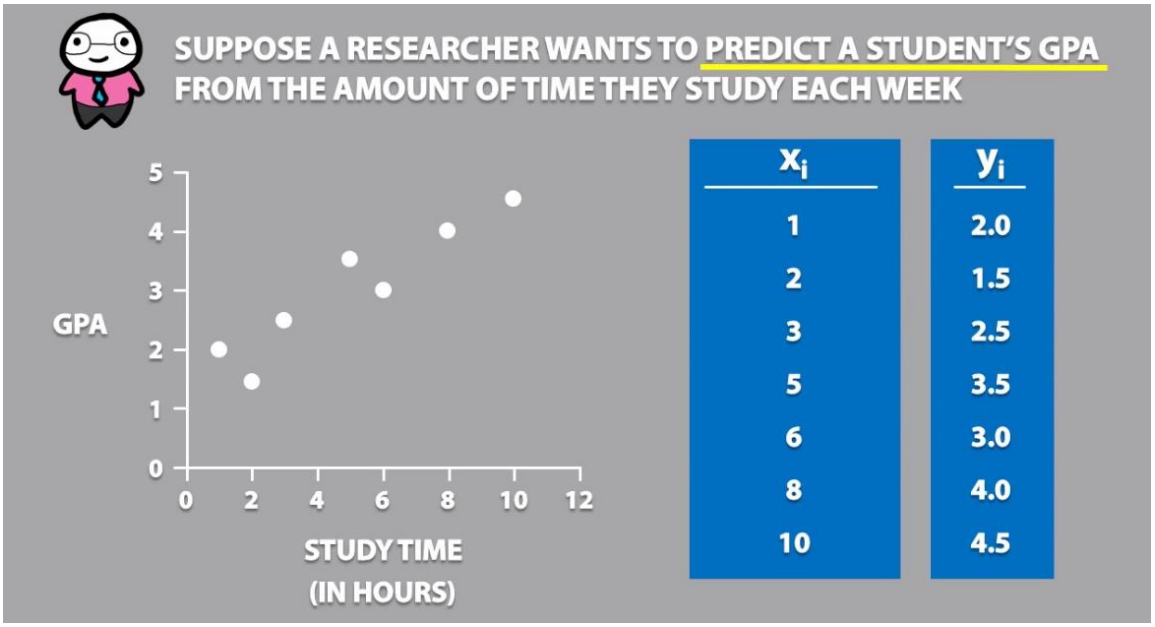

SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA
FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK
STUDENT STUDY TIME GPA
GEORGE 1 2.0
VANESSA 2 1.5
GATSBY 3 2.5
SOPHIA 5 3.5
EMMA 6 3.0
MELISSA 8 4.0
PATRICK 10 4.5

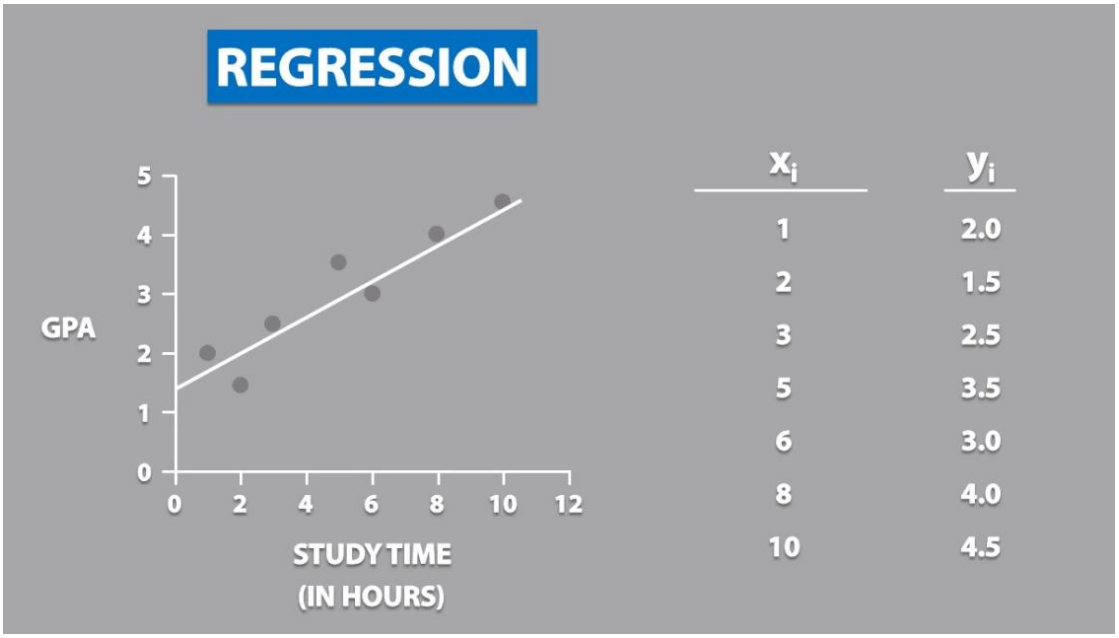

SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA
FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK
STUDY TIME
GPA
1
2.0
2
1.5
3
2.5
5
3.5
6
3.0
8
4.0
10
4.5


SUPPOSE A RESEARCHER WANTS TO PREDICT A STUDENT'S GPA
FROM THE AMOUNT OF TIME THEY STUDY EACH WEEK
GPA
5
4
3
2
1
0
STUDY TIME
y_i
1
2.0
2
1.5
3
2.5
5
3.5
6
3.0
8
4.0
10
4.5

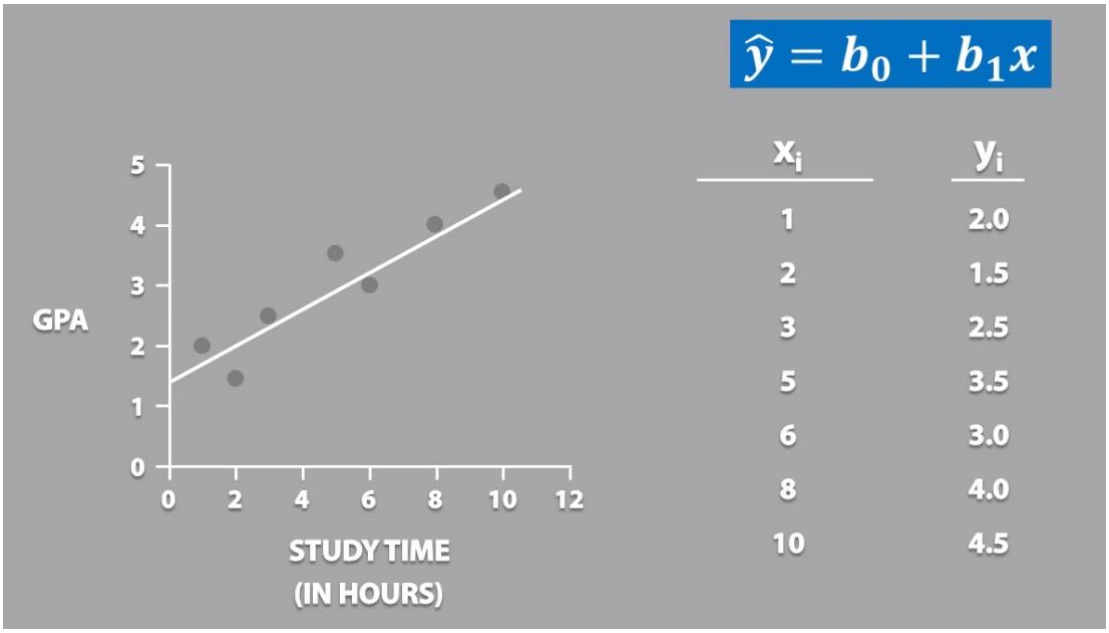


✓ Let's use scatter plot to plot this data

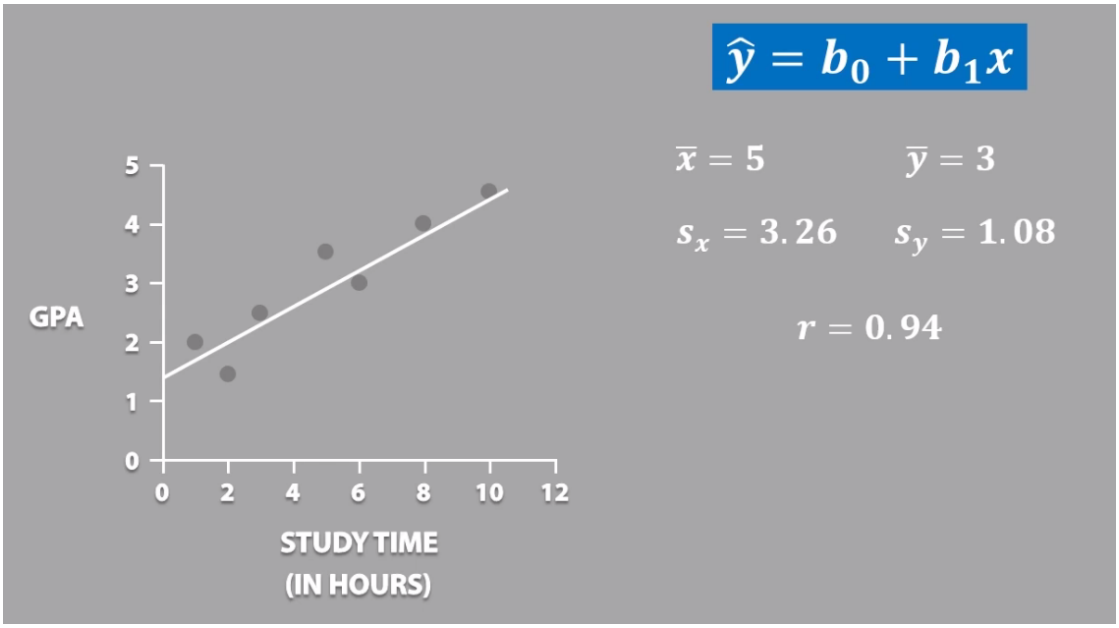


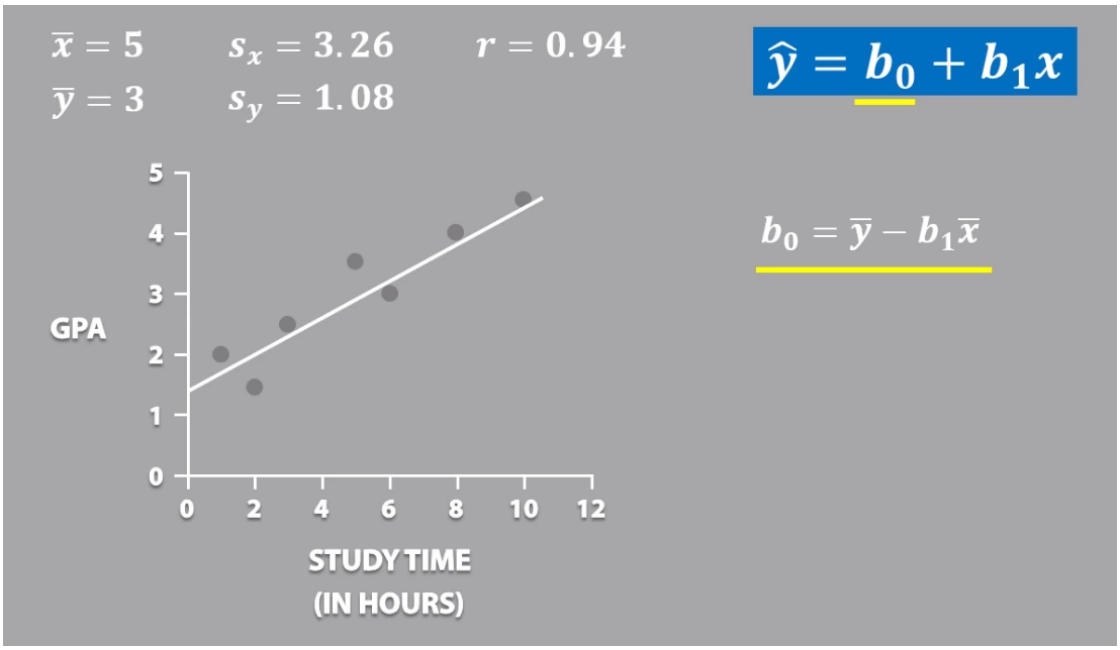


x_i	y_i
1	2.0
2	1.5
3	2.5
5	3.5
6	3.0
8	4.0
10	4.5



x_i	y_i
1	2.0
2	1.5
3	2.5
5	3.5
6	3.0
8	4.0
10	4.5

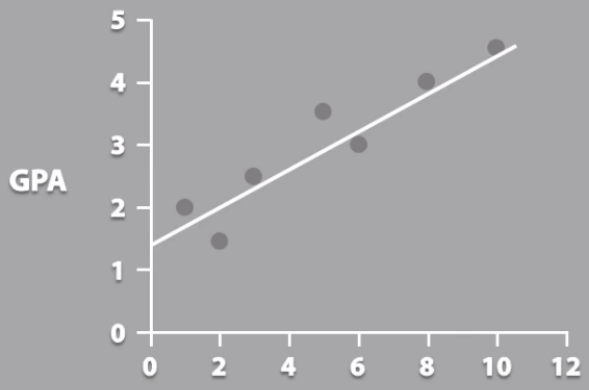

$\hat{y} = b_0 + b_1 x$
$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)


$\bar{x} = 5$
$\bar{y} = 3$
$s_x = 3.26$
$s_y = 1.08$
$r = 0.94$
$\hat{y} = b_0 + b_1 x$
$b_0 = \bar{y} - b_1 \bar{x}$
GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)

$\bar{x} = 5$ $s_x = 3.26$ $r = 0.94$

$\bar{y} = 3$ $s_y = 1.08$

$$\hat{y} = b_0 + b_1 x$$



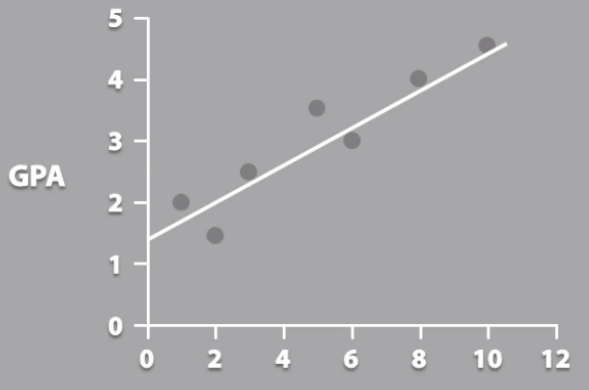
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = r \frac{s_y}{s_x}$$

$\bar{x} = 5$ $s_x = 3.26$ $r = 0.94$

$\bar{y} = 3$ $s_y = 1.08$

$$\hat{y} = b_0 + b_1 x$$



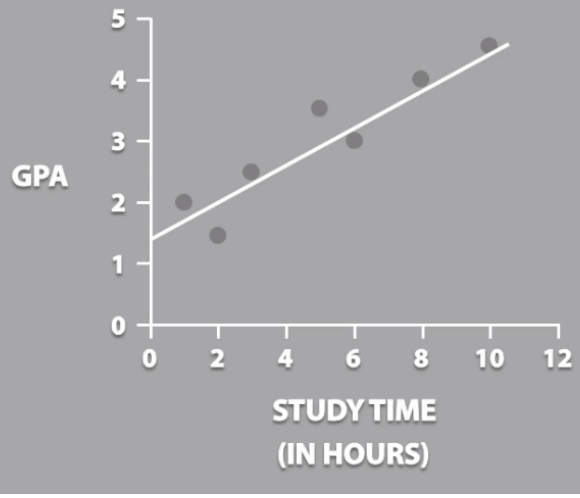
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\Rightarrow b_1 = 0.311$$

$\bar{x} = 5$ $s_x = 3.26$ $r = 0.94$

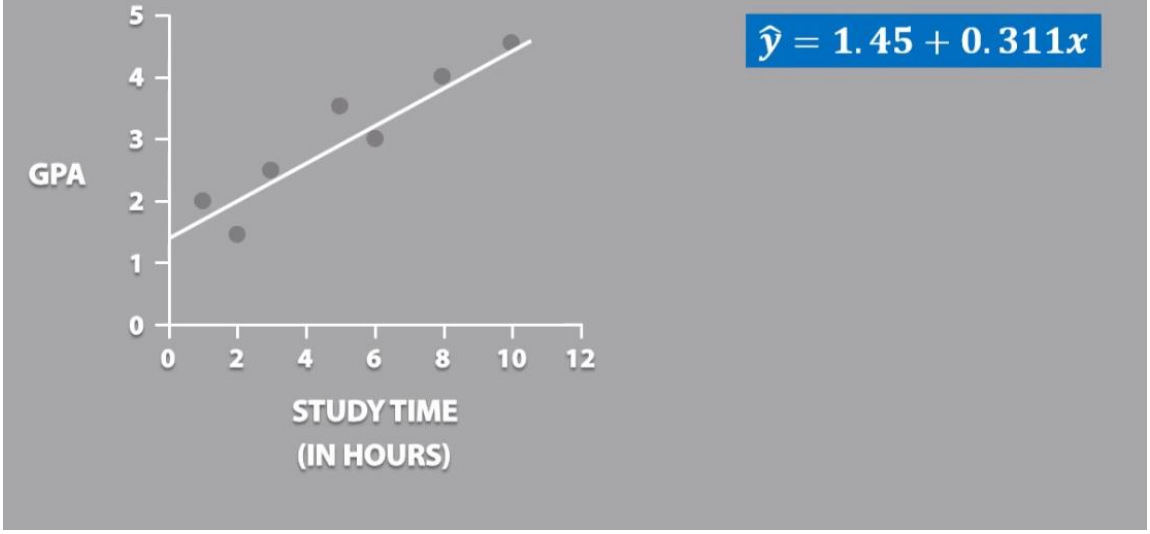
$\bar{y} = 3$ $s_y = 1.08$

$$\hat{y} = b_0 + b_1 x$$

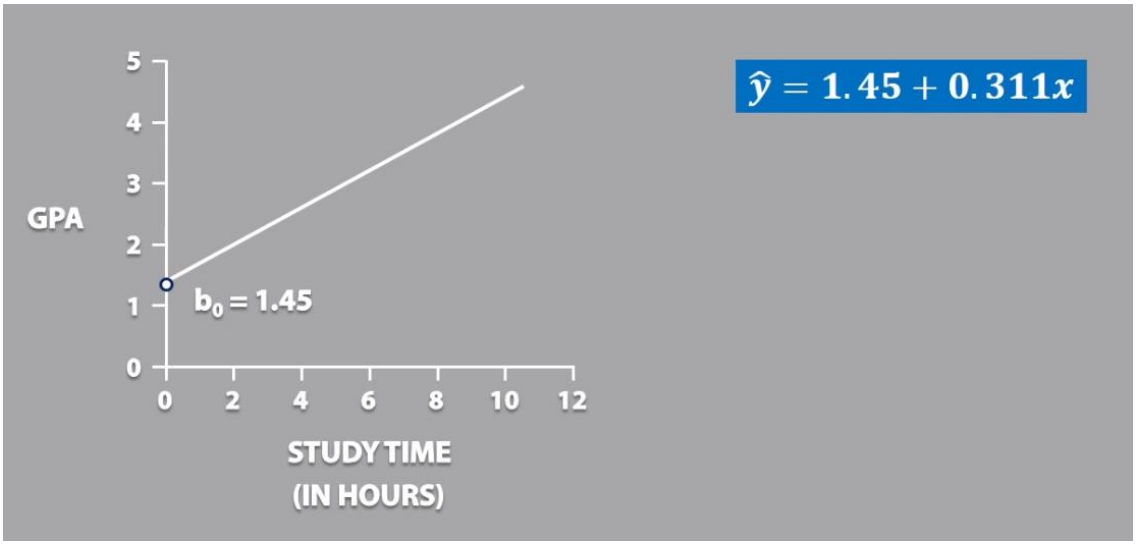


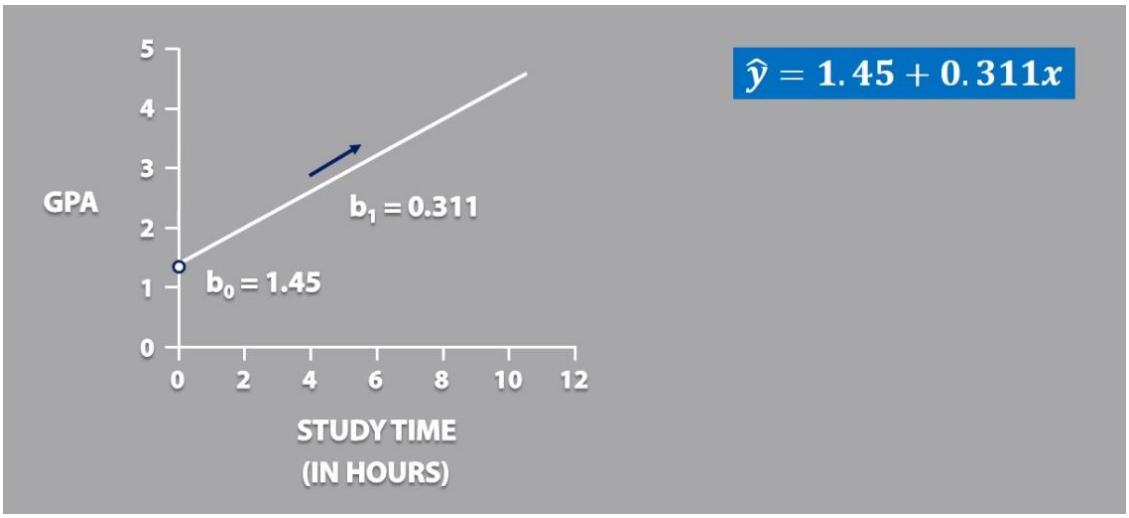
$b_0 = 1.45$

$b_1 = 0.311$



$$\hat{y} = 1.45 + 0.311x$$

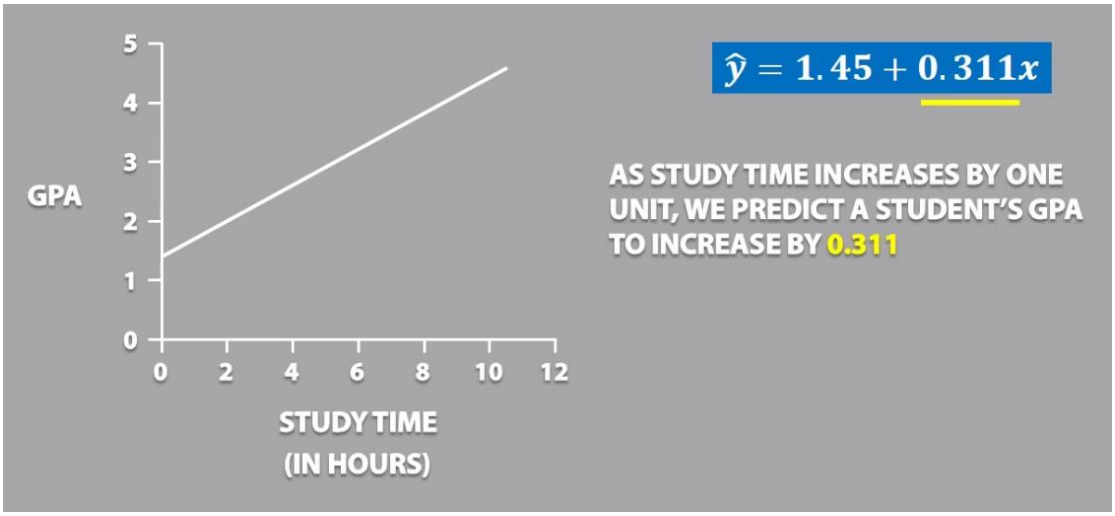

$\hat{y} = 1.45 + 0.311x$
GPA
5
4
3
2
1
0
$b_0 = 1.45$
0 2 4 6 8 10 12
STUDY TIME
(IN HOURS)

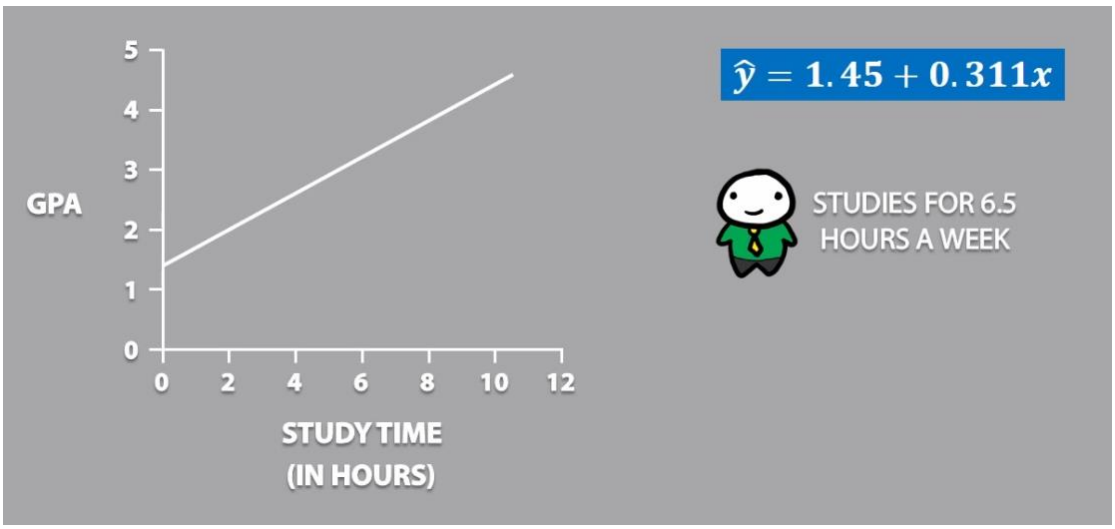

$\hat{y} = 1.45 + 0.311x$
GPA
5
4
3
2
1
0
$b_1 = 0.311$
$b_0 = 1.45$
0 2 4 6 8 10 12
STUDY TIME
(IN HOURS)

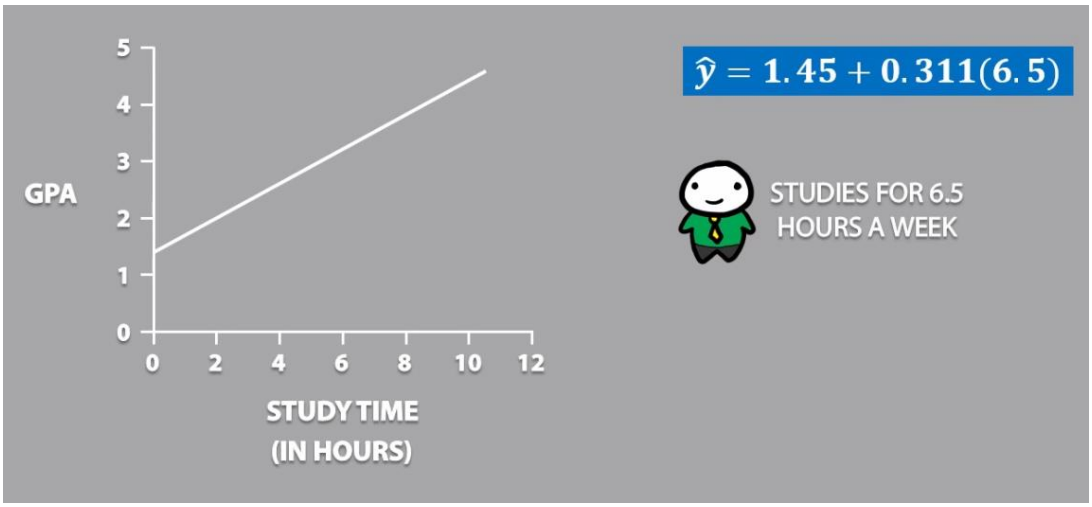
THE LINE OF LEAST SQUARES REGRESSION

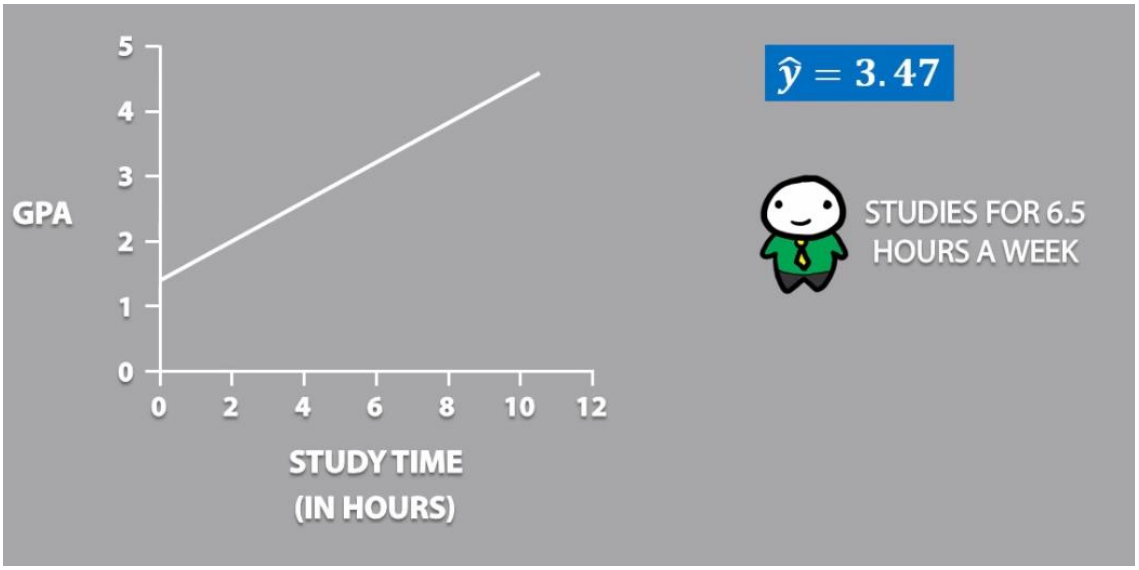
$$\hat{y} = b_0 + b_1 x$$

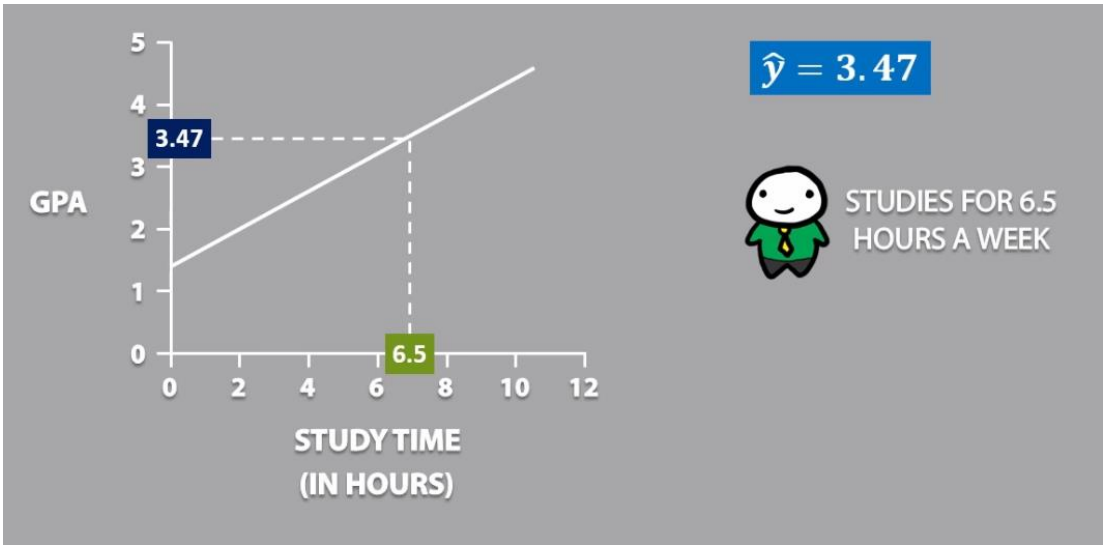
THE SLOPE OF A REGRESSION LINE PREDICTS THE CHANGE IN "Y" WHEN "X" INCREASES BY ONE UNIT



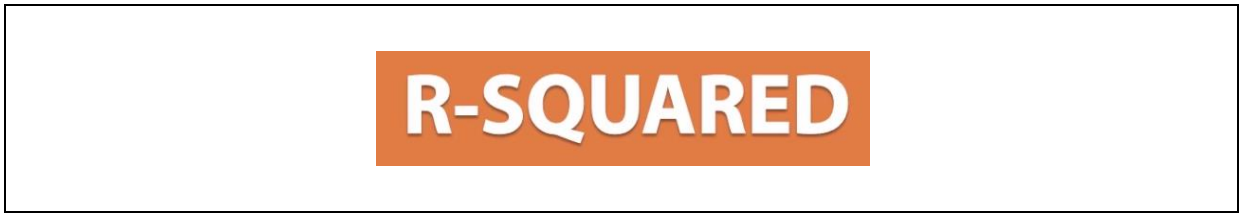

GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 1.45 + 0.311x$
STUDIES FOR 6.5
HOURS A WEEK

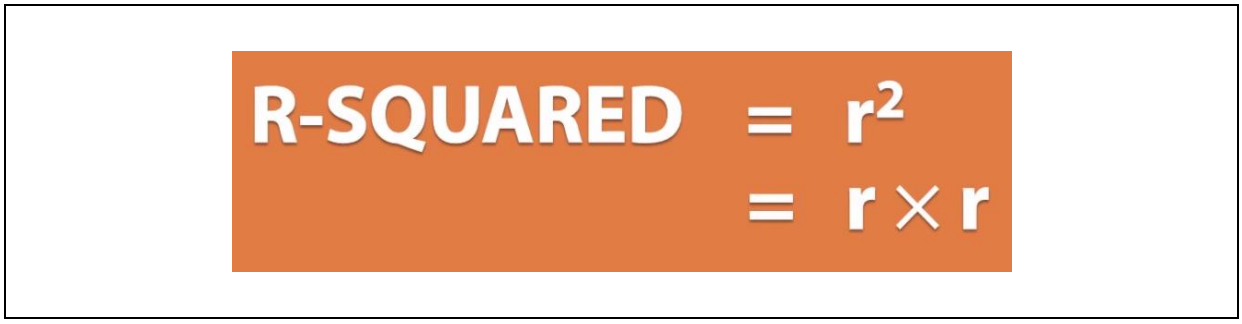

GPA
0
1
2
3
4
5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 1.45 + 0.311(6.5)$
STUDIES FOR 6.5
HOURS A WEEK


GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 3.47$
STUDIES FOR 6.5
HOURS A WEEK

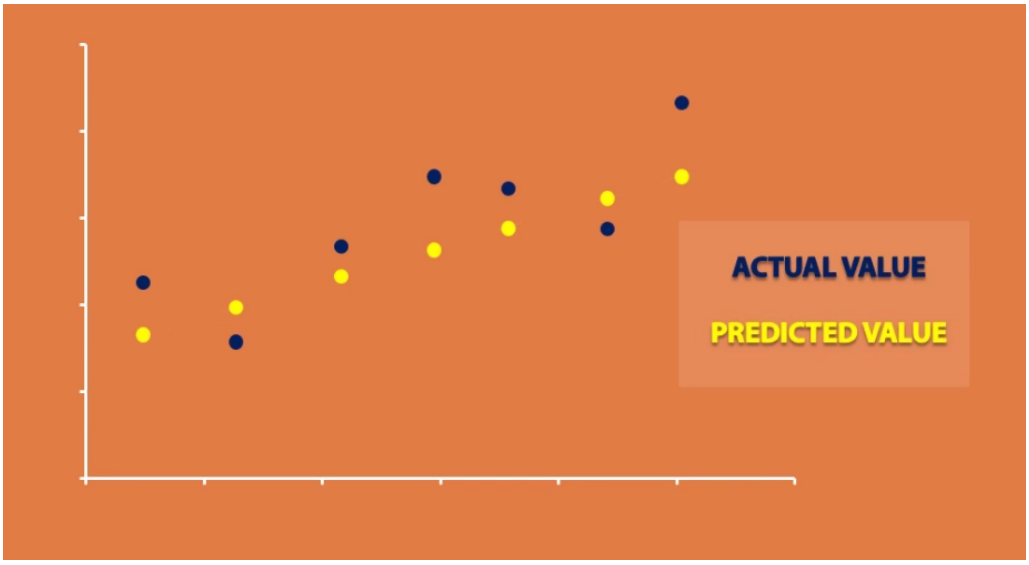

GPA
5
4
3.47
3
2
1
0
6.5
0
2
4
6
8
10
12
STUDY TIME
(IN HOURS)
$\hat{y} = 3.47$
STUDIES FOR 6.5
HOURS A WEEK

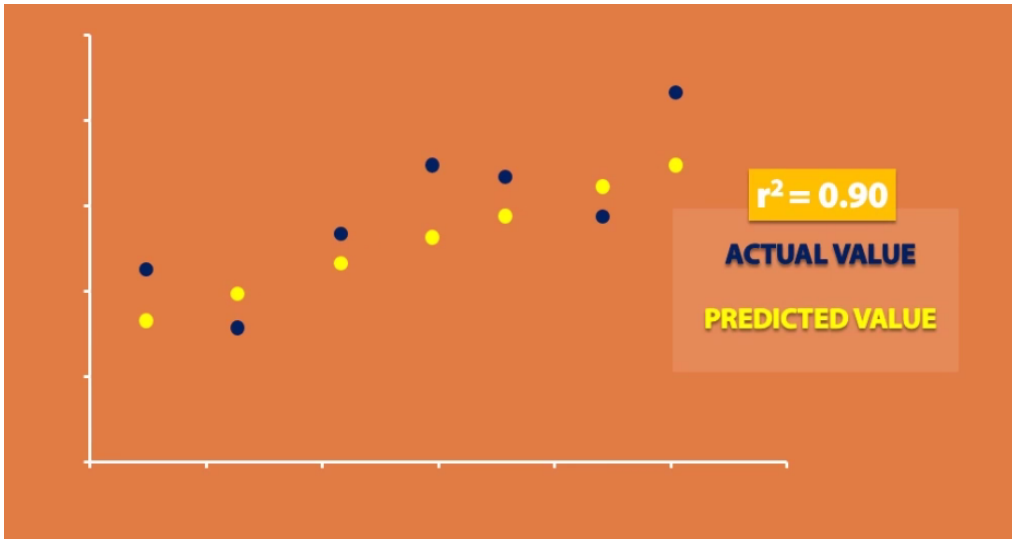
5. R – SQUARED





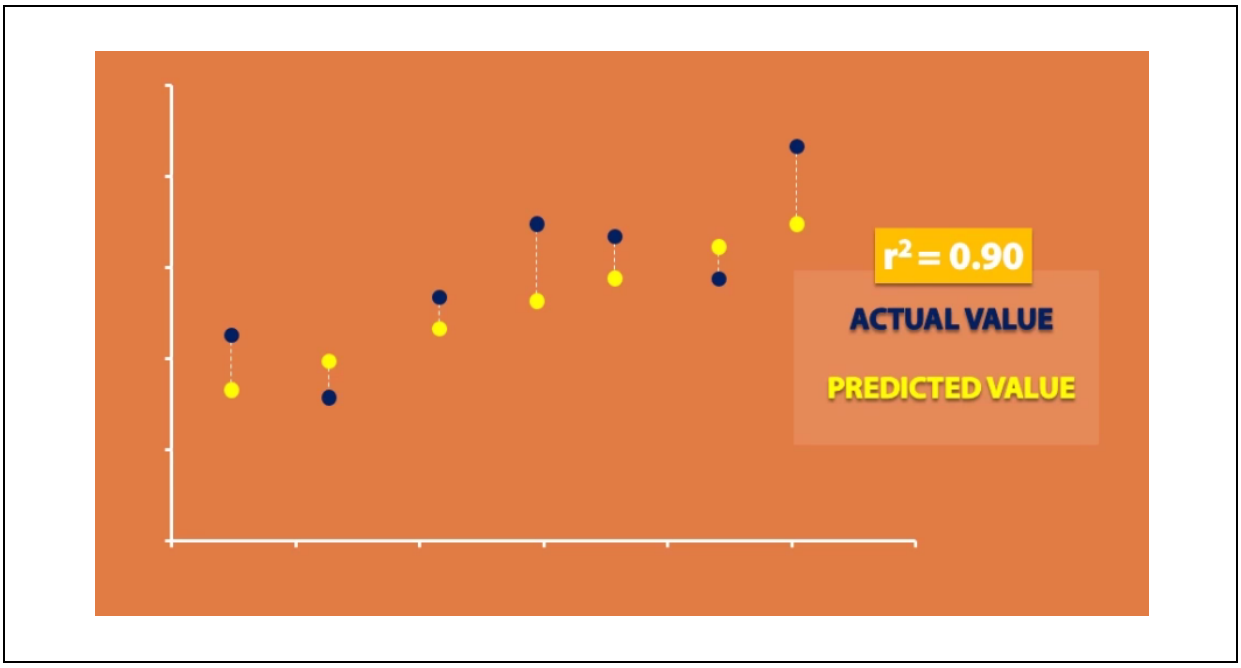
r	r^2
HAS VALUES BETWEEN -1 AND 1 MEASURES THE LINEAR RELATIONSHIP BETWEEN TWO QUANTITATIVE VARIABLES WITH RESPECT TO DIRECTION AND STRENGTH	HAS VALUES BETWEEN 0 AND 1 IS A MEASURE OF HOW CLOSE EACH DATA POINT FITS TO THE REGRESSION LINE TELLS US HOW WELL THE REGRESSION LINE PREDICTS ACTUAL VALUES




r² = 0.90
ACTUAL VALUE
PREDICTED VALUE

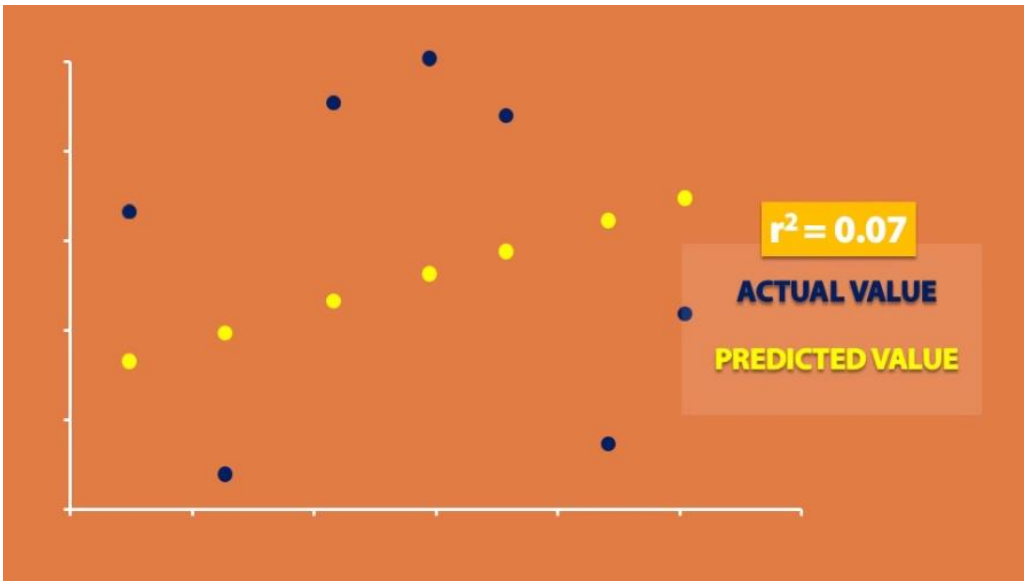
Case 1: r squared value is high

- ✓ A high value of r square value explains the actual values and predicted values are close together
- ✓ We can clear see the actual values are predicted values having very less distance in between them



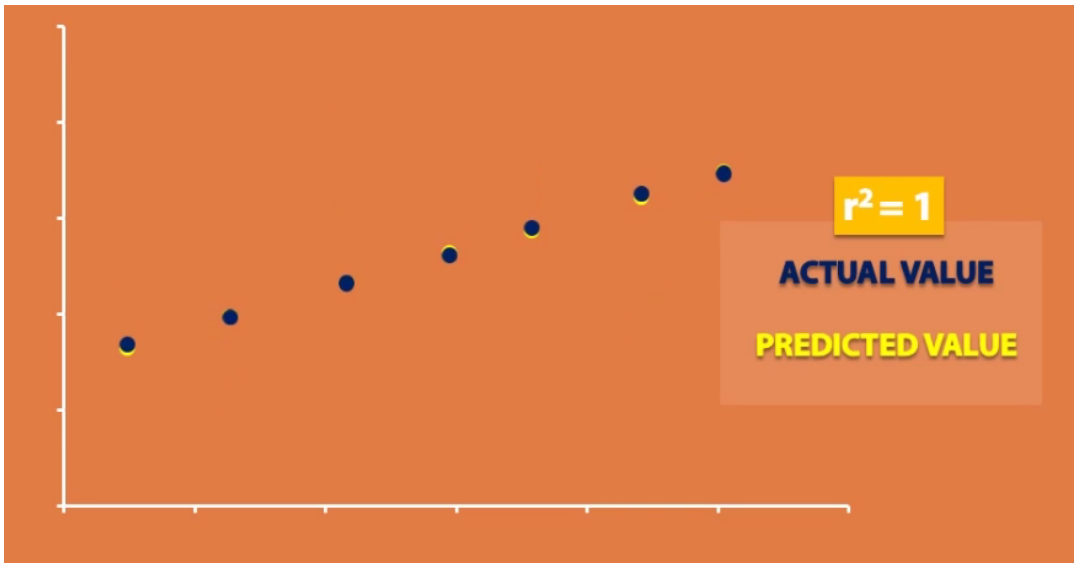
Case 2: r squared value is low

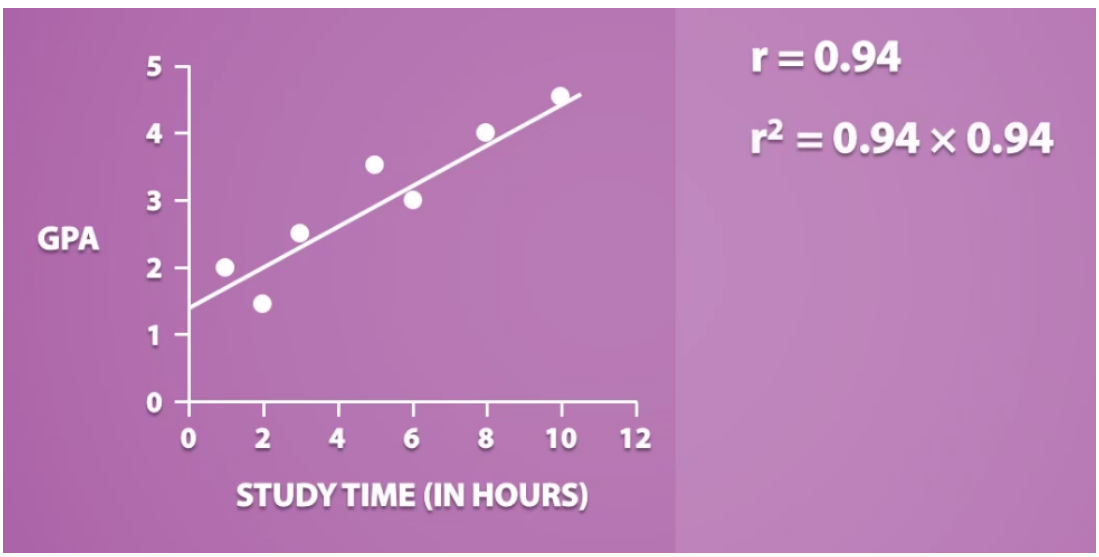
- ✓ A low value of r square value explains the regression line which the data points are does not fit the well
- ✓ We can clear see the actual values are predicted values having large distance in between them

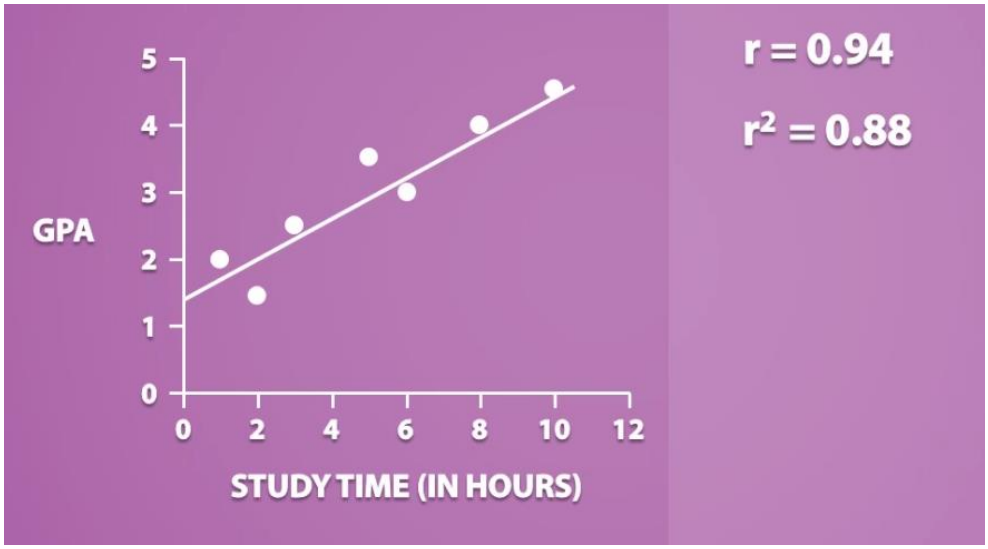


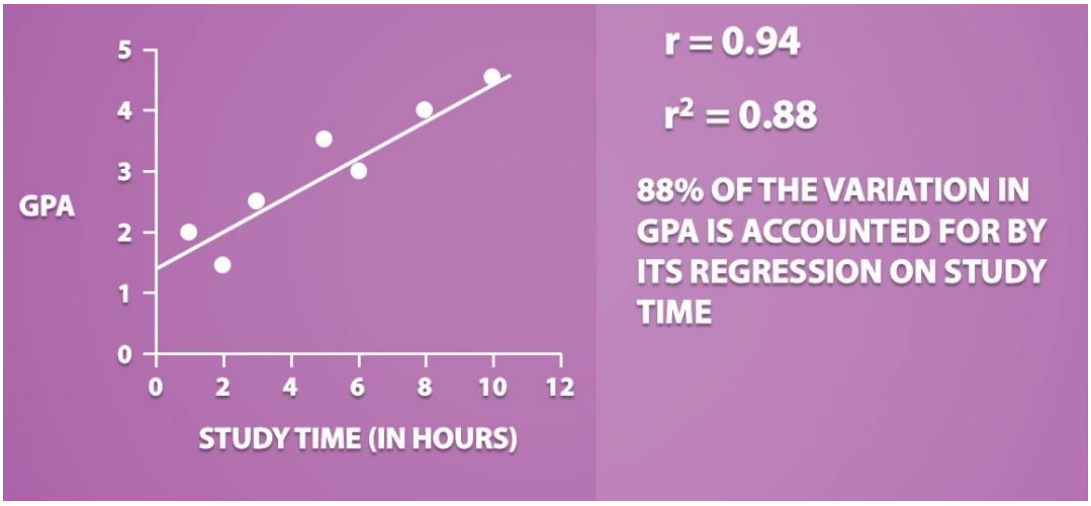
Case 3: r squared value is equals to one

- ✓ This means that we can predict value of y for any given value of x






GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)
r = 0.94
r² = 0.88


GPA
5
4
3
2
1
0
0
2
4
6
8
10
12
STUDY TIME (IN HOURS)
r = 0.94
r² = 0.88
88% OF THE VARIATION IN GPA IS ACCOUNTED FOR BY ITS REGRESSION ON STUDY TIME