

Linear Correlation and Regression

Notes

Bivariate Data

An ordered pair of data values (x,y) representing two quantities.

X is known as the independent variable, which can be controlled.

y is known as the dependent variable, which can't be controlled.

Why?

We gather bivariate data so that we can determine whether there is a linear relationship between the quantities x and y .

Specifically, we want to determine whether a linear correlation exists between two variables x and y .

What do you mean
by Linear?

$$y = mx + b$$

Algebra

The following sample of data values represent the amount of time studied (hours) and the corresponding grade (percent) earned for 6 students.

X	y
Hours Studied	Grade(Percent)
0	5
2	24
4	43
6	62
8	81
10	100

Algebra

Using the points (0,5) and (4,43)
we have the following.

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{43 - 5}{4 - 0} = 9.5$$

$$y - y_1 = m(x - x_1)$$

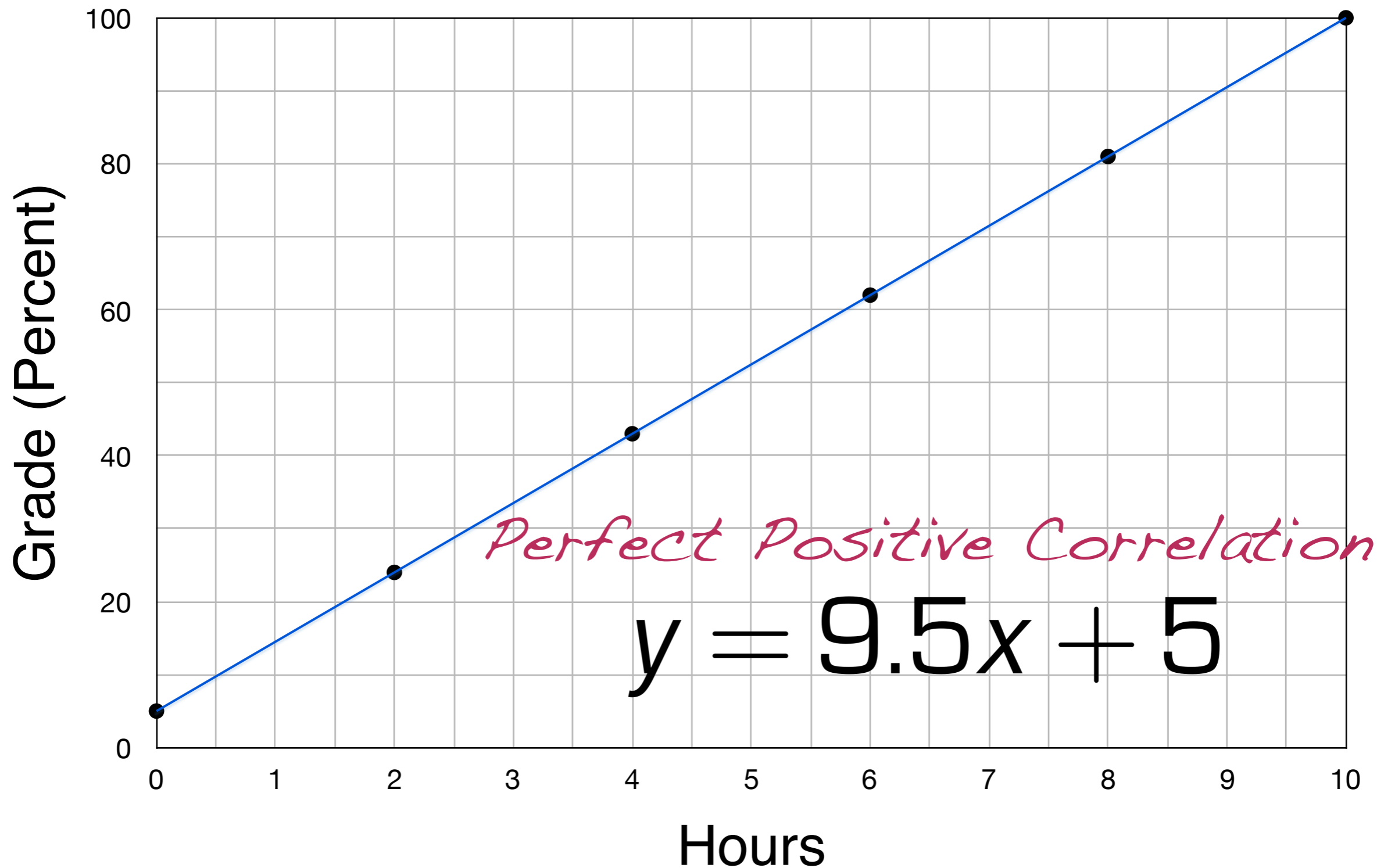
$$y - 5 = 9.5(x - 0)$$

$$y - 5 = 9.5x$$

$$y = 9.5x + 5$$

Scatter Plot

Study Time Versus Score



Why is it important to have a correlation?

2 We can use the equation that describes the data to predict!

$$y = 9.5x + 5$$

Prediction

If you study for 3 hours, what grade will you earn?

$$y = 9.5x + 5$$

$$y = 9.5 \cdot 3 + 5$$

$$y = 33.5$$

Prediction

If you study for 5 hours, what grade will you earn?

$$y = 9.5x + 5$$

$$y = 9.5 \cdot 5 + 5$$

$$y = 52.5$$

Prediction

If you study for 7 hours, what grade will you earn?

$$y = 9.5x + 5$$

$$y = 9.5 \cdot 7 + 5$$

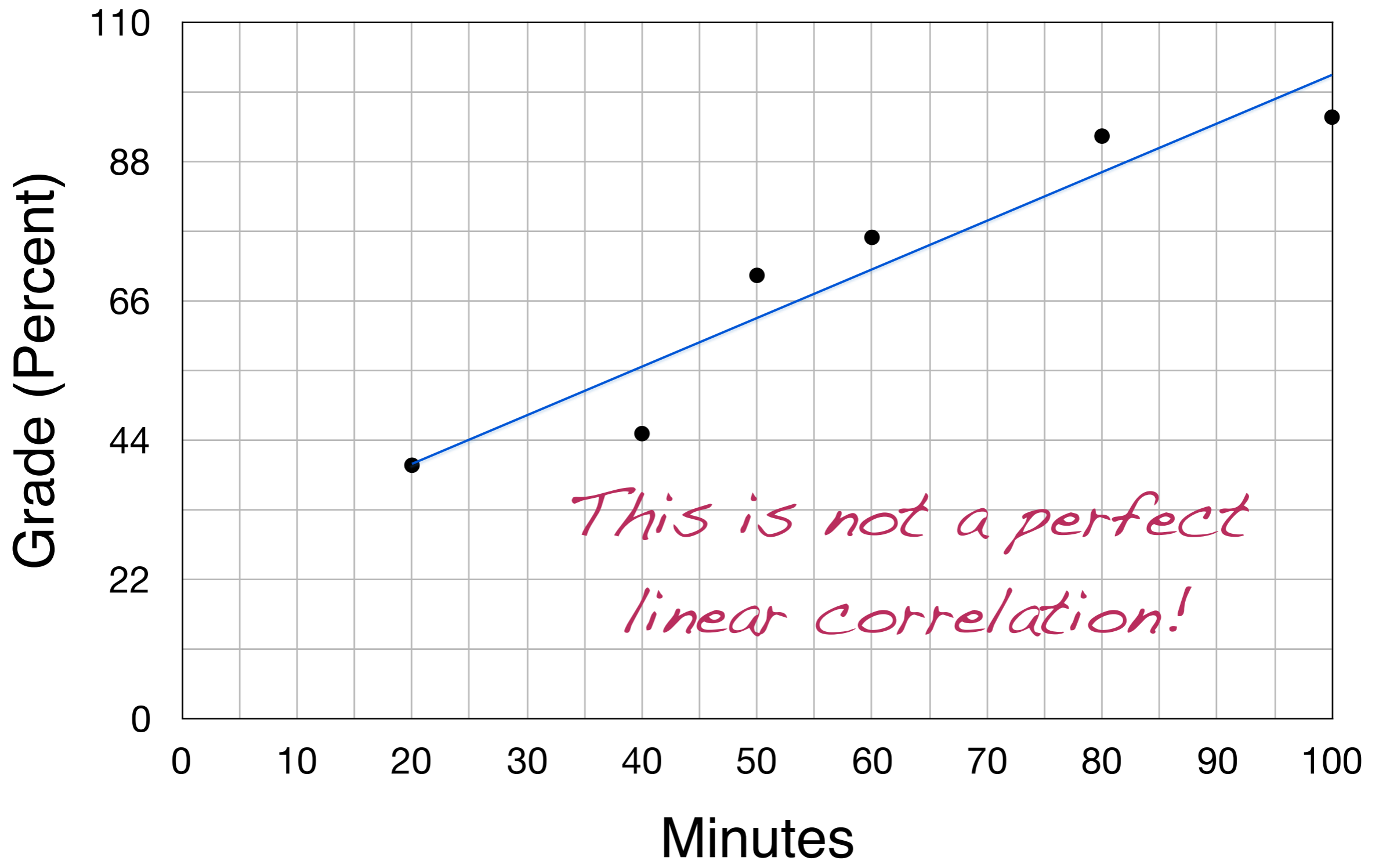
$$y = 71.5$$

The following data values represent the amount of time studied (minutes) and the corresponding grade (percent) earned.

X		y
Min Studied		Grade (Percent)
20		40
40		45
50	<i>Realistic</i>	70
60		76
80		92
100		95

Scatter Plot

Study Time Versus Score



÷ Can we still use a
linear equation to

Maybe?

2
We will have to determine whether there is a *linear correlation* between the variables x and y . And, if there is a linear correlation, we can determine the *linear equation* that “Best Fits” our data.

Linear Correlation Coefficient

r

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

Determine the linear correlation coefficient r .

X Min Studied	y Grade(Percent)
20	40
40	45
50	70
60	76
80	92
100	95

Note the sample size is 6.

	A	B	C	D	E	F	G
1		x	y	xy	x^2	y^2	
2		20	40				
3		40	45				
4		50	70				
5		60	76				
6		80	92				
7		100	95				
8							
9	Sum						
10							
11							
12							
13							
14							
15							
16							

Level of Significance

α

used to identify the cutoff (critical value) between results attributed to chance and results attributed to an actual relationship between the two variables.

Using the 5% level of significance

$$|r| > \textit{critical value}$$

$$|0.95| > 0.811$$

$$0.95 > 0.811$$

The data gathered and assuming no linear correlation between x and y, there's a 5% chance

$$|r| > \textit{critical value}$$

$$|0.95| > 0.811$$

$$0.95 > 0.811$$

It's unusual for r to satisfy this condition.
Therefore, we have a linear correlation

The “Best Fit Line”

$$\hat{y} = mx + b$$

where

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

	A	B	C	D	E	F	G
1		x	y	xy	x^2	y^2	
2		20	40	800	400	1600	
3		40	45	1800	1600	2025	
4		50	70	3500	2500	4900	
5		60	76	4560	3600	5776	
6		80	92	7360	6400	8464	
7		100	95	9500	10000	9025	
8							
9	Sum	350	418	27520	24500	31790	
10							
11	m						
12	b						
13							
14							
15							
16							

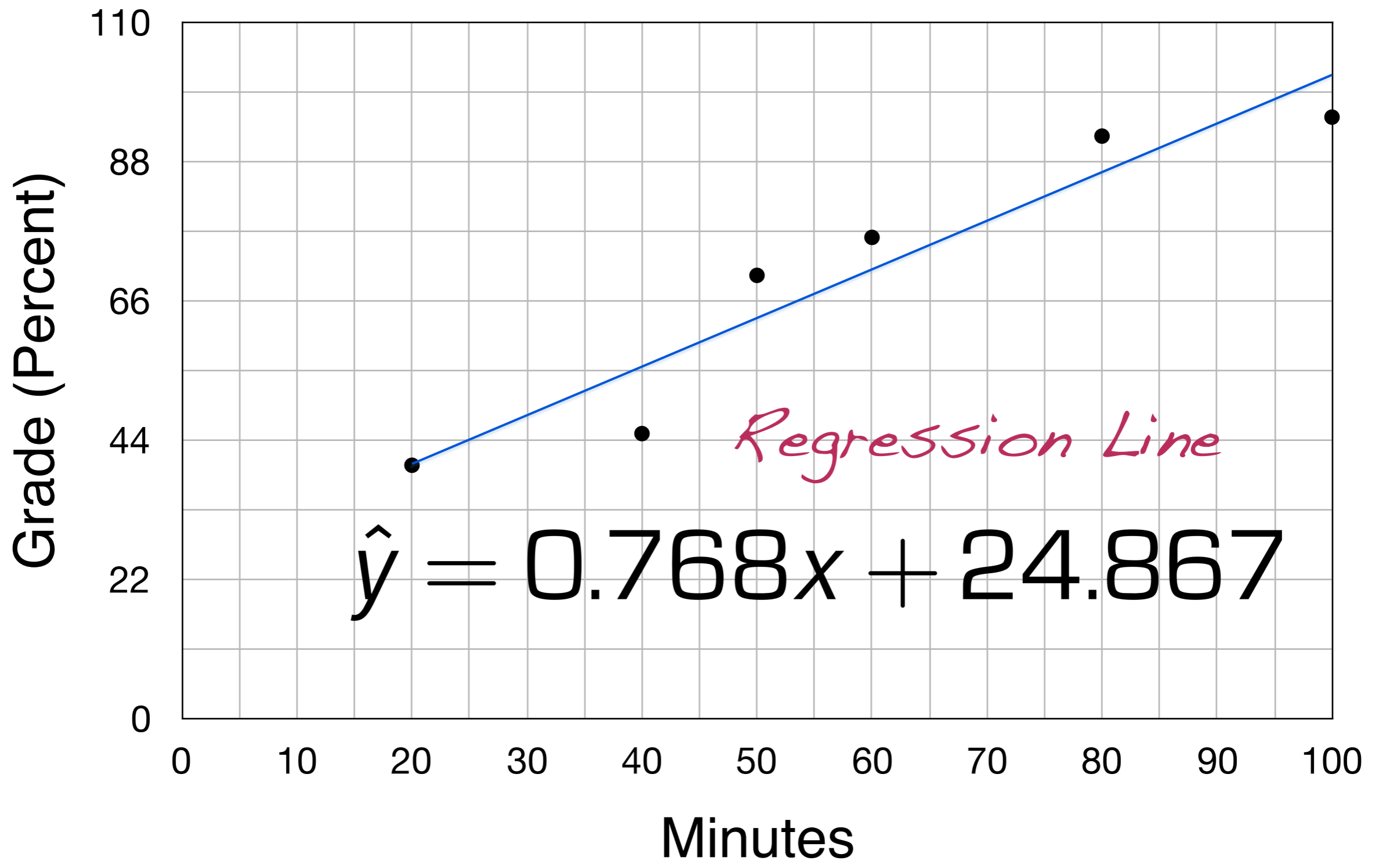
The “Best Fit Line”

$$\hat{y} = 0.7688x + 24.8667$$

aka, the Regression Line

Scatter Plot

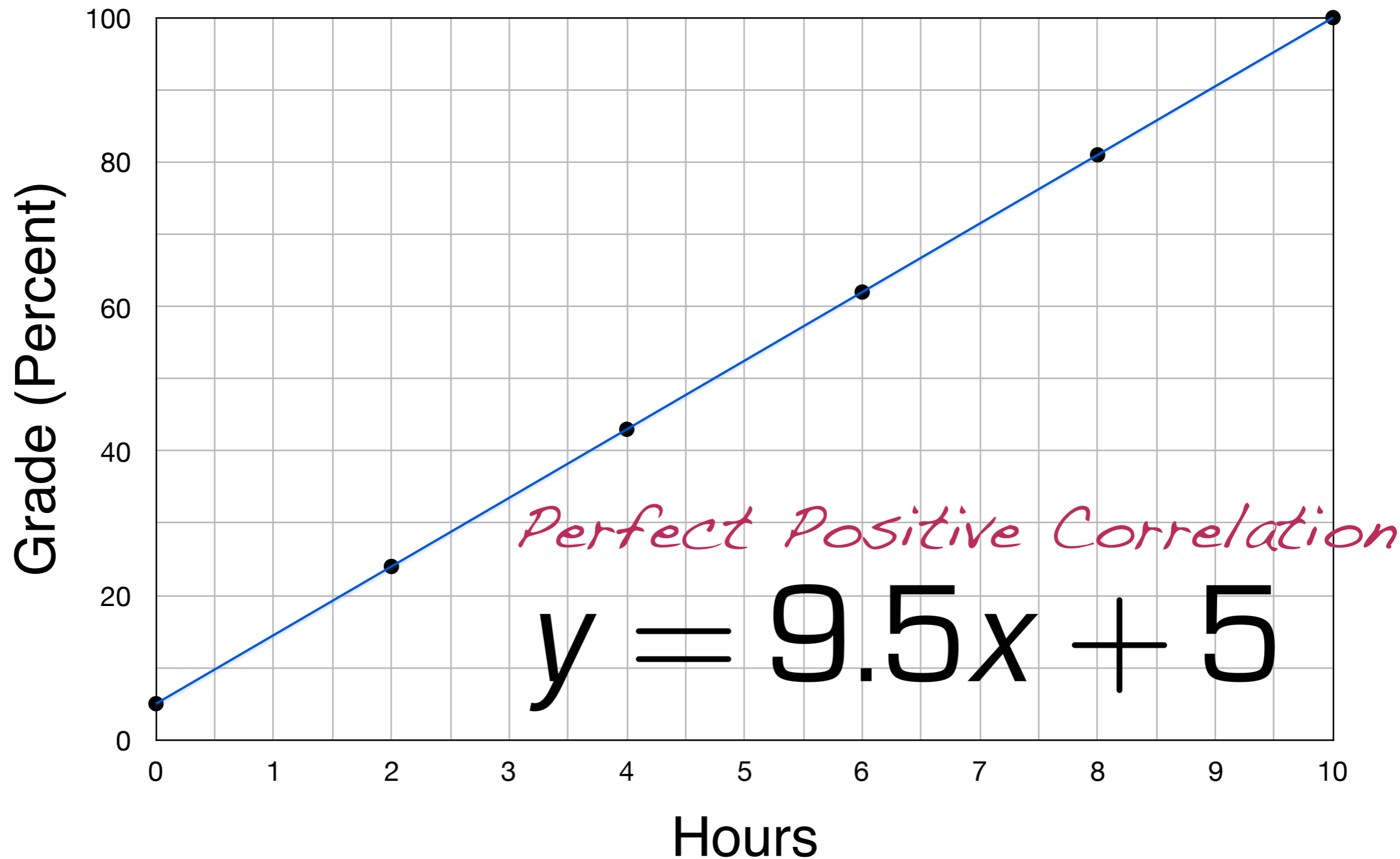
Study Time Versus Score



Recall

Scatter Plot

Study Time Versus Score



Determine the linear correlation coefficient r .

x	y
Hours Studied	Grade(Percent)
0	5
2	24
4	43
6	62
8	81
10	100

Note the sample size is 6.

	A	B	C	D	E	F	G
1		x	y	xy	x^2	y^2	
2		0	5				
3		2	24				
4		4	43				
5		6	62				
6		8	81				
7		10	100				
8							
9	Sum						
10							
11	r						
12	m						
13	b						
14							
15							
16							

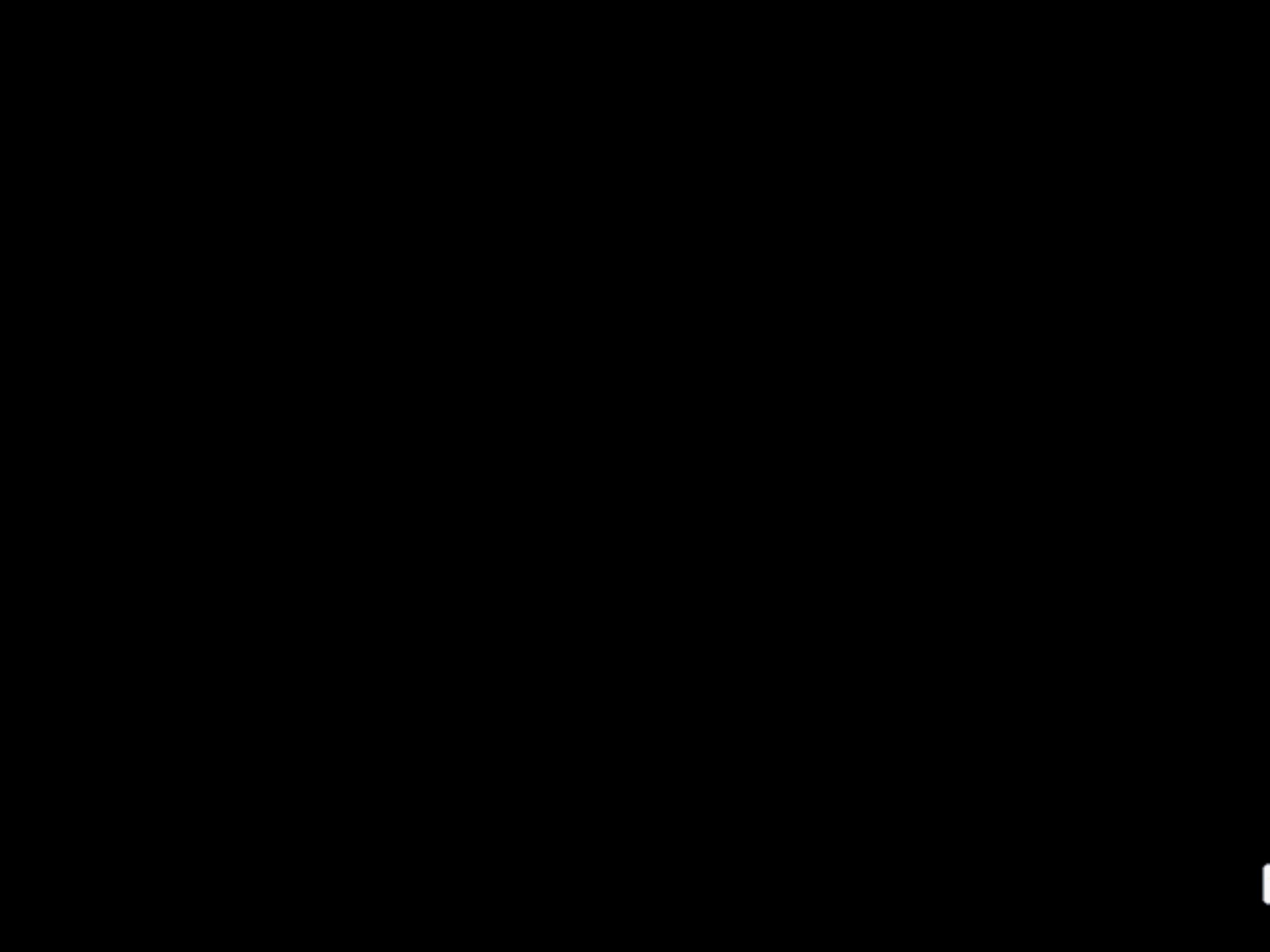
The “Best Fit Line”

$$y = 9.5x + 5$$

With the linear correlation
coefficient

$$r=1$$

*Scatter Plots and
Correlation*



Is there a linear correlation between the number of long distance calls made (monthly) and the cost (dollars)?

Determine whether a linear correlation exists at the 1% level of significance. If so, determine the “Best Fit Line”.

X # of calls	y Cost (Dollars)
10	37
15	43
12	37
20	49
25	54
17	45

Note the sample size is 6.

	A	B	C	D	E	F	G
1		x	y	xy	x ²	y ²	
2		10	37				
3		15	43				
4		12	37				
5		20	49				
6		25	54				
7		17	45				
8							
9	Sum						
10							
11	r						
12	m						
13	b						
14							
15							
16							

The data gathered and assuming no linear correlation between x and y, there's a 1% chance

$$|r| > \textit{critical value}$$

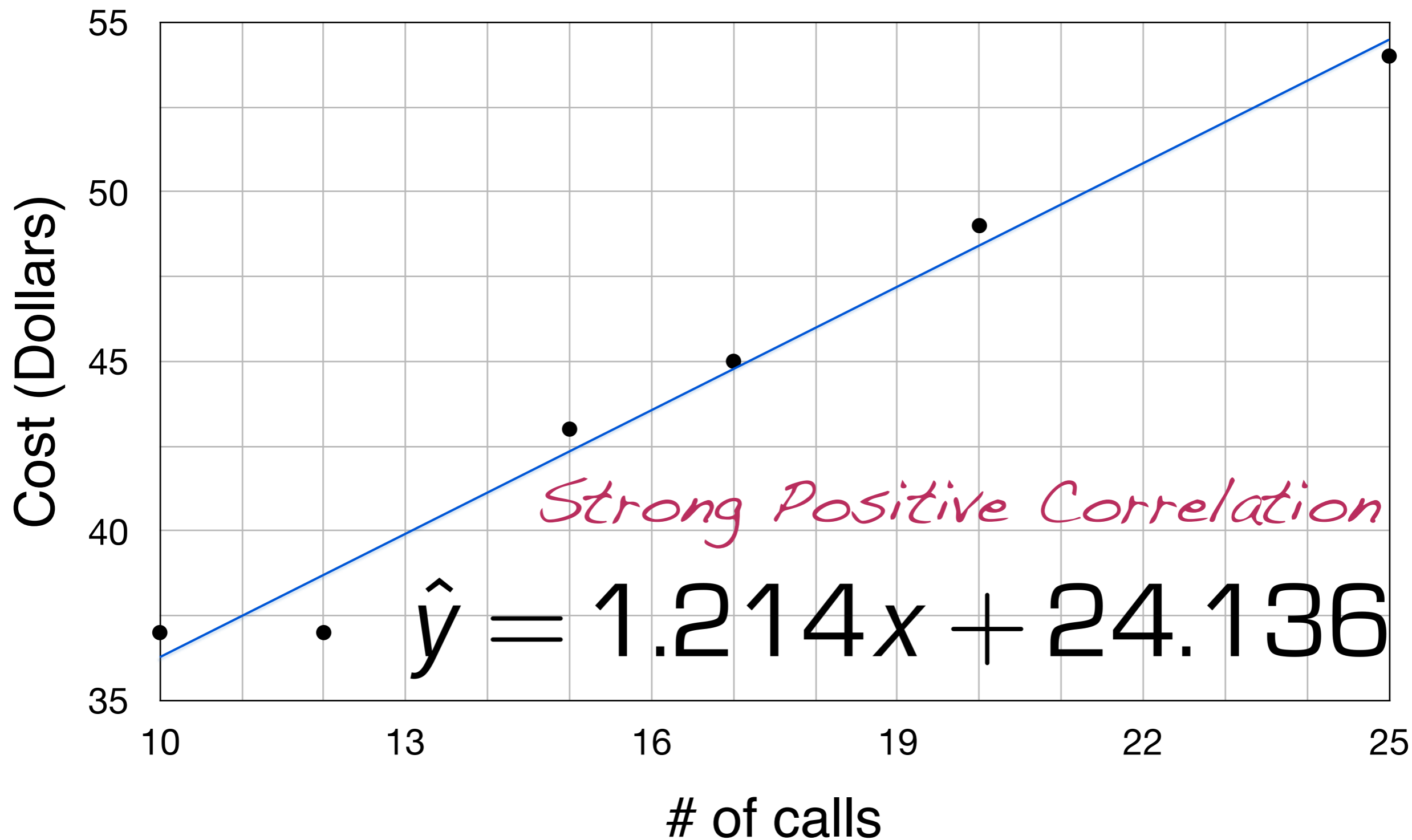
$$|0.99| > 0.917$$

$$0.99 > 0.917$$

It's unusual for r to satisfy this condition.
Therefore, we have a linear correlation

Scatter Plot

Calls versus Cost



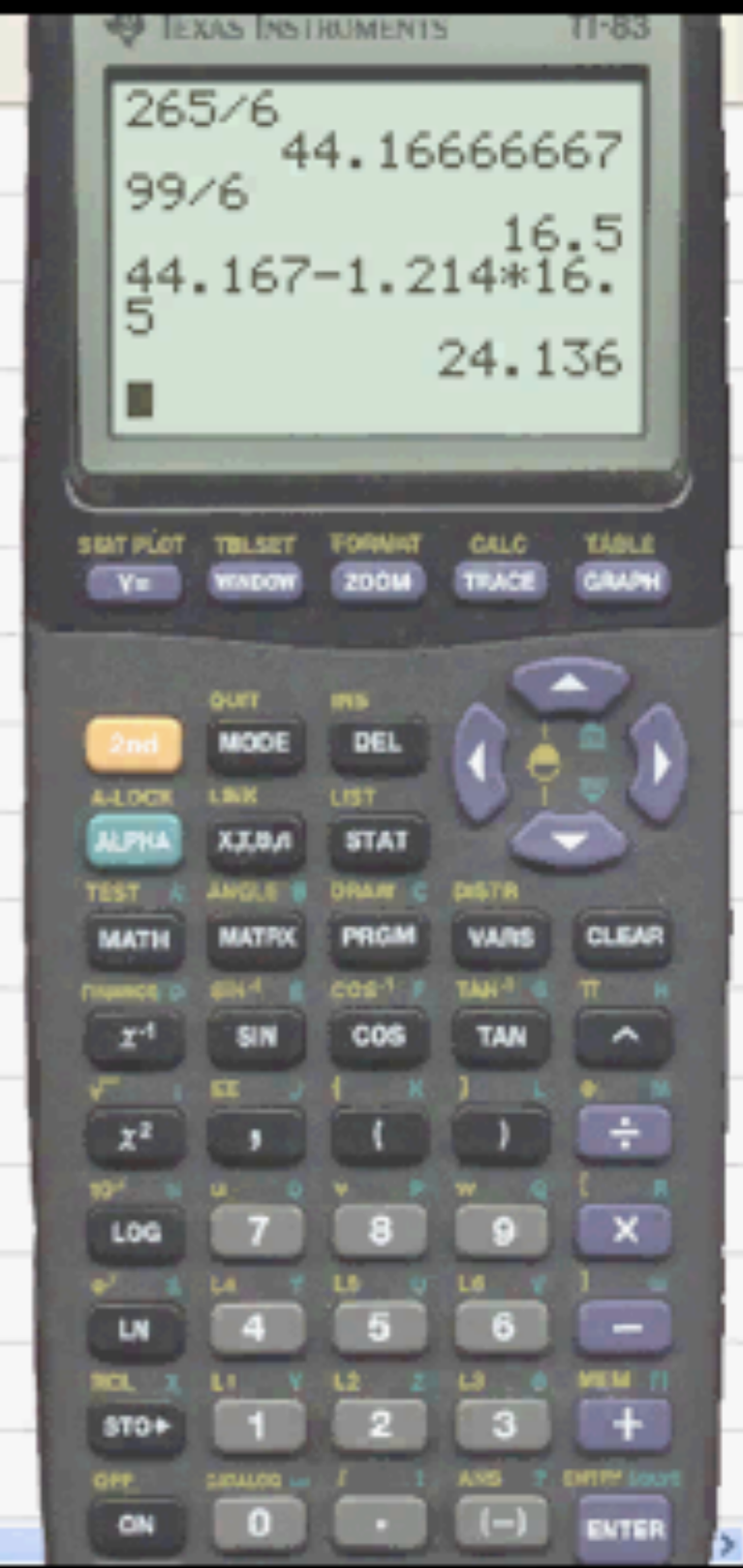
Is there a linear correlation between the number of police officers patrolling an area and the number of muggings?

Determine whether a linear correlation exists at the 5% level of significance. If so, determine the “Best Fit Line”.

X	y
Police Officers	Muggings
20	8
12	10
18	12
15	9
22	6
10	15
20	7
12	18

Note the sample size is 8.

	A	B	C	D	E
1		x	y	xy	x²
2		10	37	370	100
3		15	43	645	225
4		12	37	444	144
5		20	49	980	400
6		25	54	1350	625
7		17	45	765	289
8					
9	Sum	99	265	4554	1783
10					
11	r	0.99			
12	m	1.214			
13	b	24.136			
14					
15					
16					



The data gathered and assuming no linear correlation between x and y , there's a 5% chance

$$|r| > \textit{critical value}$$

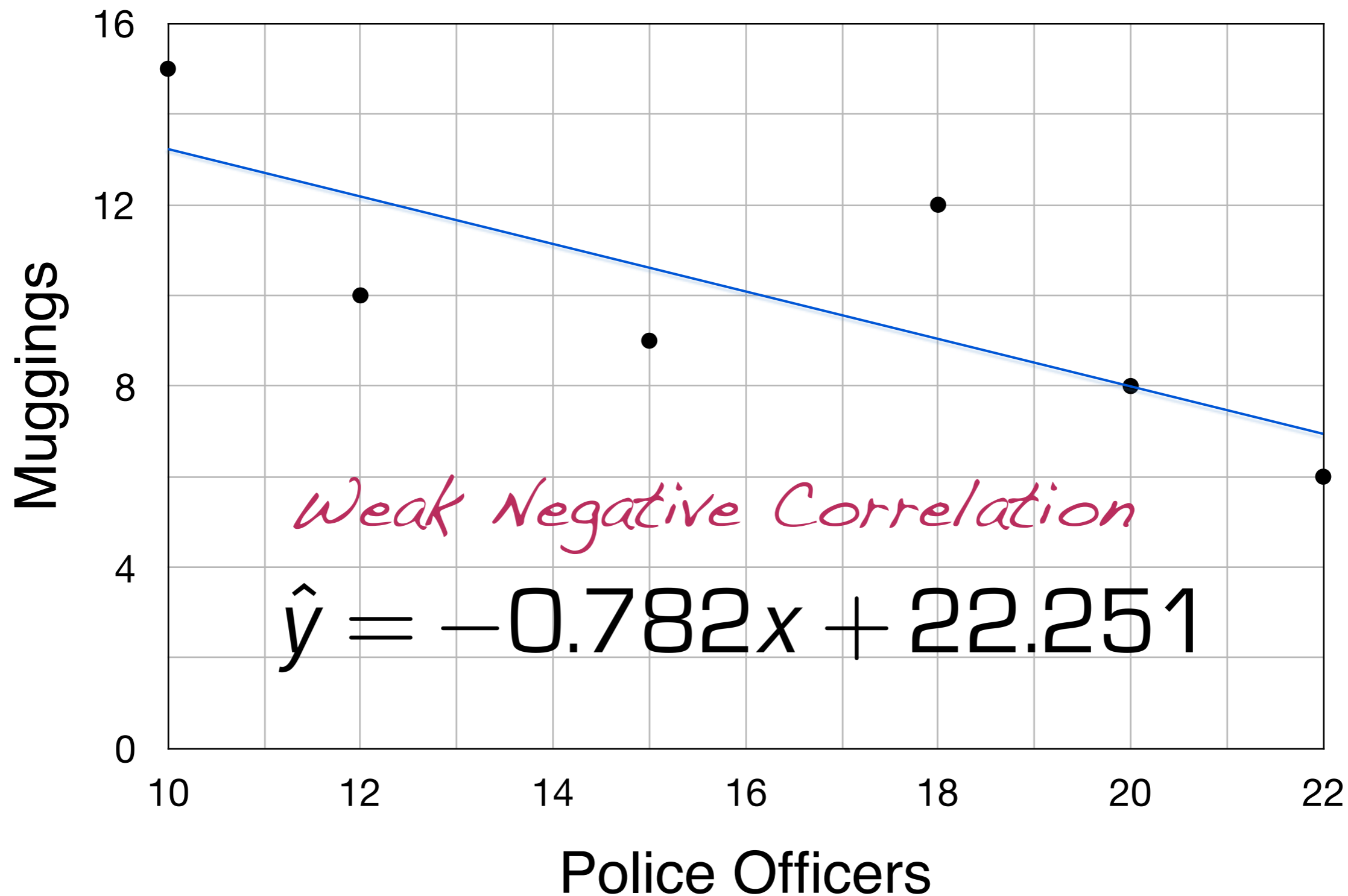
$$|-0.782| > 0.707$$

$$0.782 > 0.707$$

It's unusual for r to satisfy this condition.
Therefore, we have a linear correlation

Scatter Plot

Police Officers versus Muggings



Is there a linear correlation between the cumulative weight loss for a product over a period of weeks?

Determine whether a linear correlation exists at the 1% level of significance. If so, determine the “Best Fit Line”.

X	y
# of Weeks	Cumulative Weight Loss (pounds)
1	2
2	5
3	5
4	6
5	5
6	7
7	9

Note the sample size is 7.

	A	B	C	D	E	F	G
1		x	y	xy	x^2	y^2	
2		1	2				
3		2	5				
4		3	5				
5		4	6				
6		5	5				
7		6	7				
8		7	9				
9							
10	Sum						
11							
12	r						
13	m						
14	b						
15							
16							

The data gathered and assuming no linear correlation between x and y, there's a 1% chance

$$|r| > \textit{critical value}$$

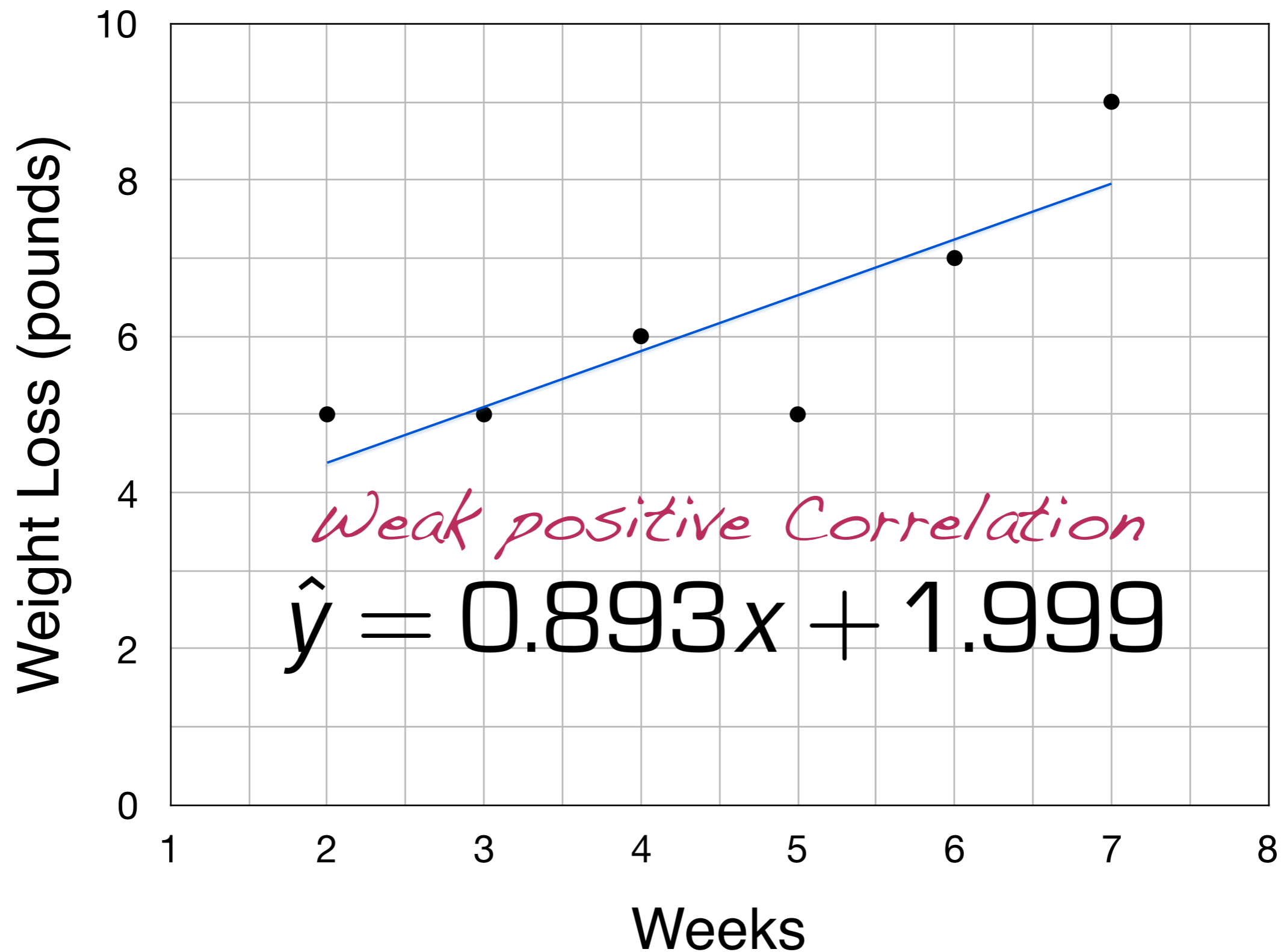
$$|0.897| > 0.875$$

$$0.897 > 0.875$$

It's unusual for r to satisfy this condition.
Therefore, we have a linear correlation

Scatter Plot

Weeks versus Weight Loss



Is there a linear correlation between the age of "runners" and the miles run per week?

Determine whether a linear correlation exists at the 1% level of significance. If so, determine the “Best Fit Line”.

X	y
Age	Miles Run (per week)
18	6
34	8
50	4
18	4
25	10
62	12
20	6

Note the sample size is 7.

	A	B	C	D	E	F	G
1		x	y	xy	x ²	y ²	
2		18	6				
3		34	8				
4		50	4				
5		18	4				
6		25	10				
7		62	12				
8		20	6				
9							
10	Sum						
11							
12	r						
13	m						
14	b						
15		+					
16							

The data gathered and assuming no linear correlation between x and y, there's a 1% chance

$$|r| > \textit{critical value}$$

$$|0.459| > 0.875$$

$$0.459 \not> 0.875$$

There is not sufficient evidence to support the conclusion of a linear correlation.

Scatter Plot

Age versus Miles Run

