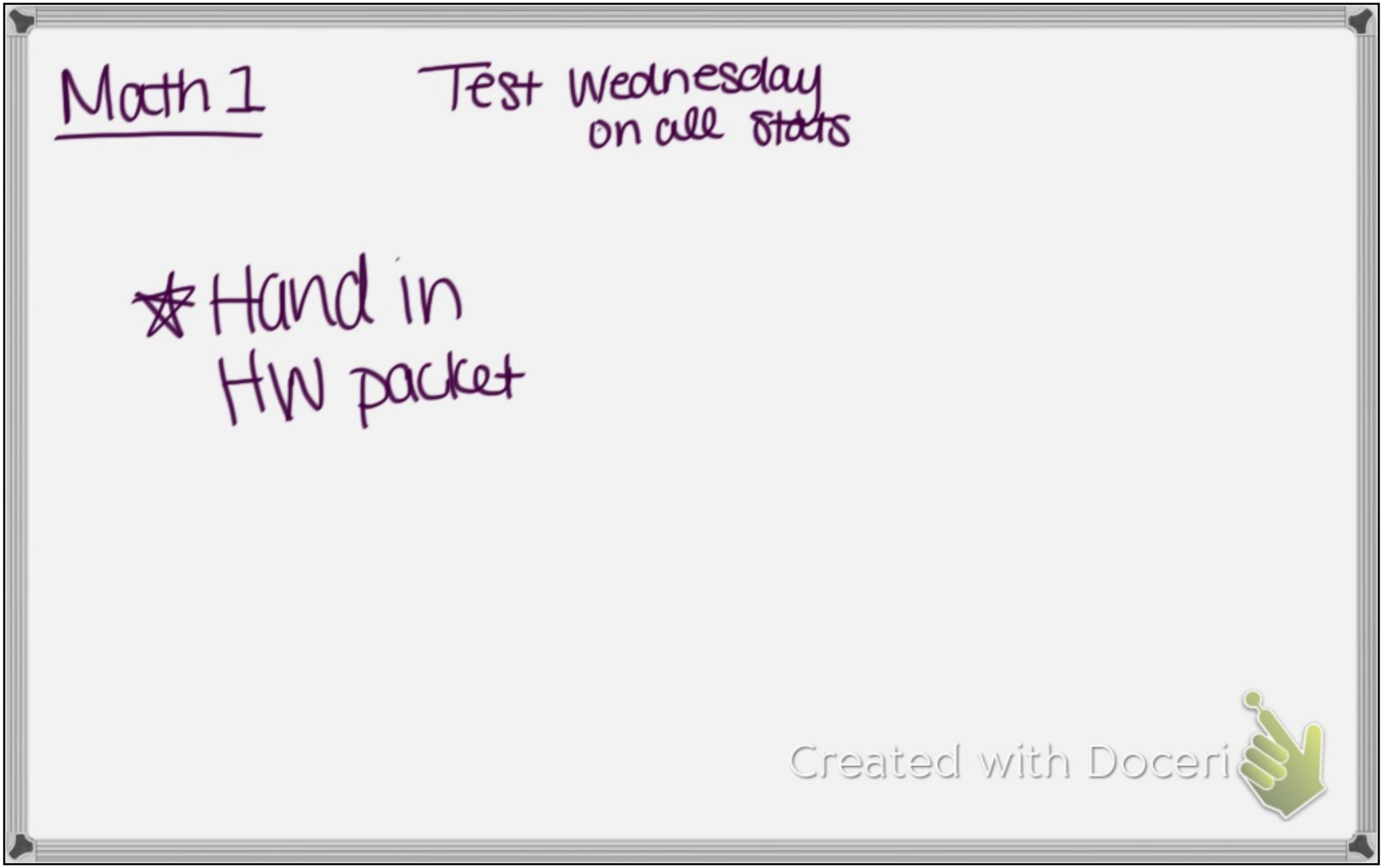
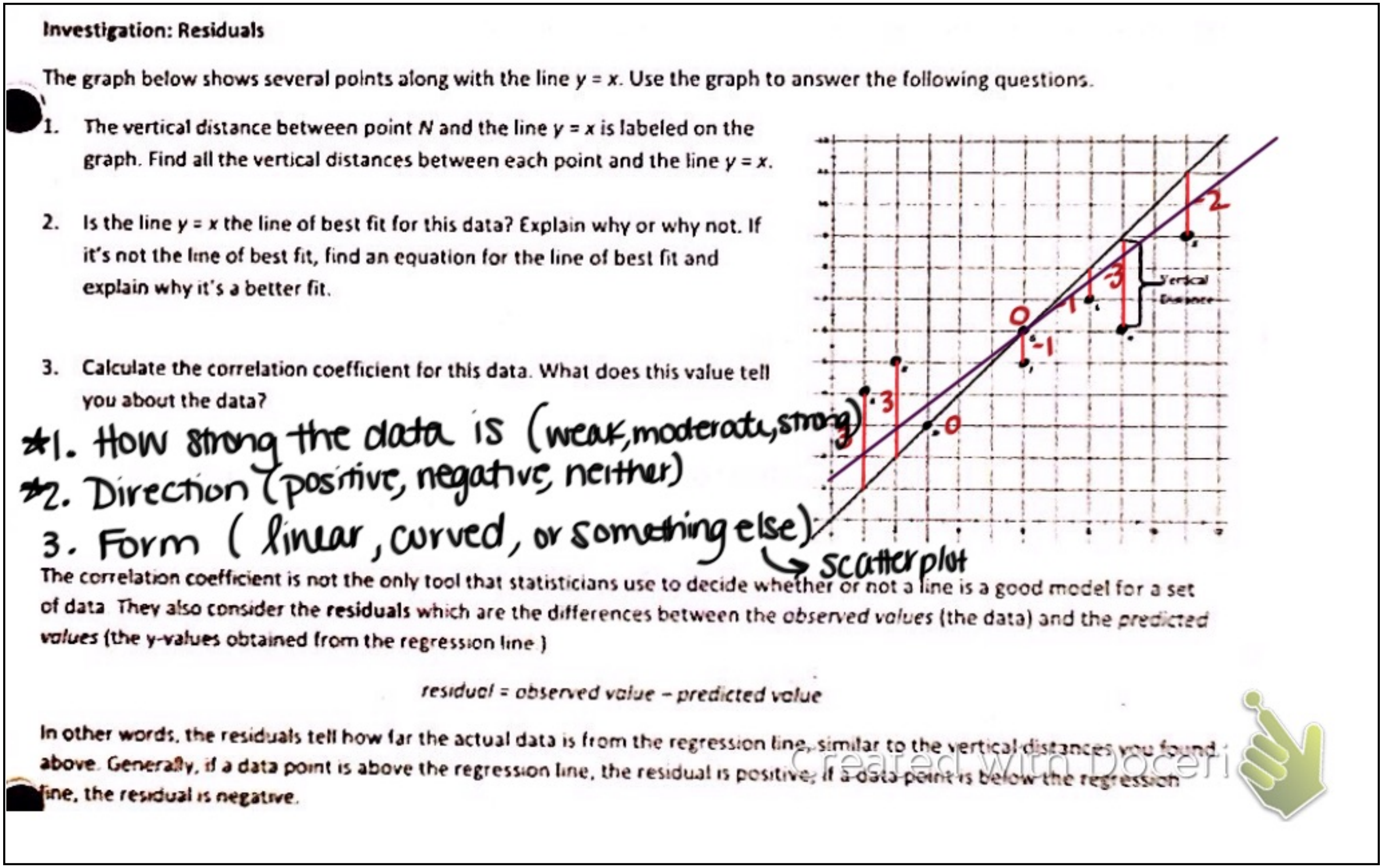


Math 1

Test Wednesday
on all stats

★ Hand in
HW packet



Investigation: Residuals

The graph below shows several points along with the line $y = x$. Use the graph to answer the following questions.

1. The vertical distance between point N and the line $y = x$ is labeled on the graph. Find all the vertical distances between each point and the line $y = x$.

2. Is the line $y = x$ the line of best fit for this data? Explain why or why not. If it's not the line of best fit, find an equation for the line of best fit and explain why it's a better fit.

3. Calculate the correlation coefficient for this data. What does this value tell you about the data?

*1. How strong the data is (weak, moderate, strong)

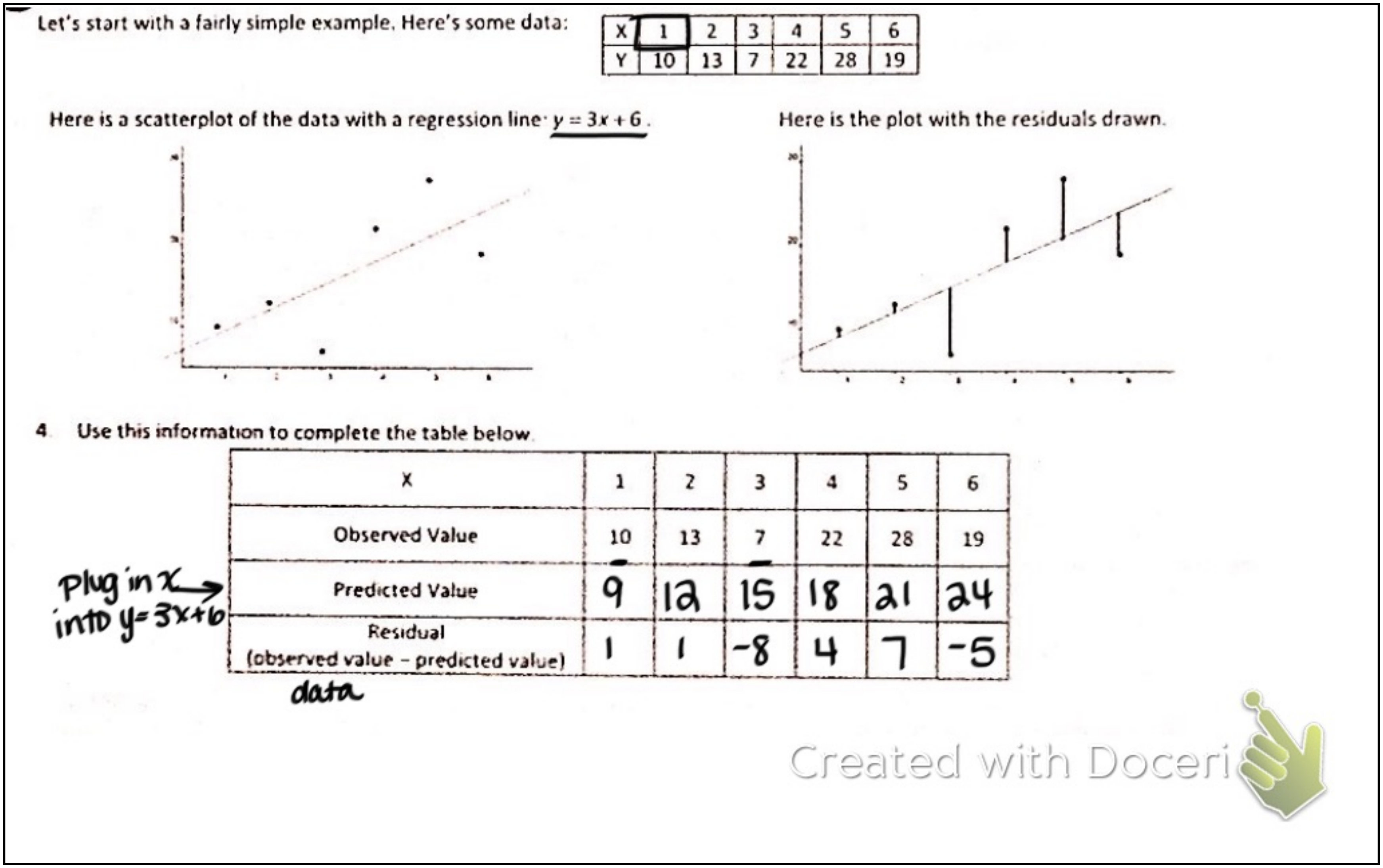
*2. Direction (positive, negative, neither)

3. Form (linear, curved, or something else) → scatterplot

The correlation coefficient is not the only tool that statisticians use to decide whether or not a line is a good model for a set of data. They also consider the **residuals** which are the differences between the *observed values* (the data) and the *predicted values* (the y-values obtained from the regression line.)

residual = observed value − predicted value

In other words, the residuals tell how far the actual data is from the regression line, similar to the vertical distances you found above. Generally, if a data point is above the regression line, the residual is positive; if a data point is below the regression line, the residual is negative.



Let's start with a fairly simple example. Here's some data:

X	1	2	3	4	5	6
Y	10	13	7	22	28	19

Here is a scatterplot of the data with a regression line: $y = 3x + 6$.

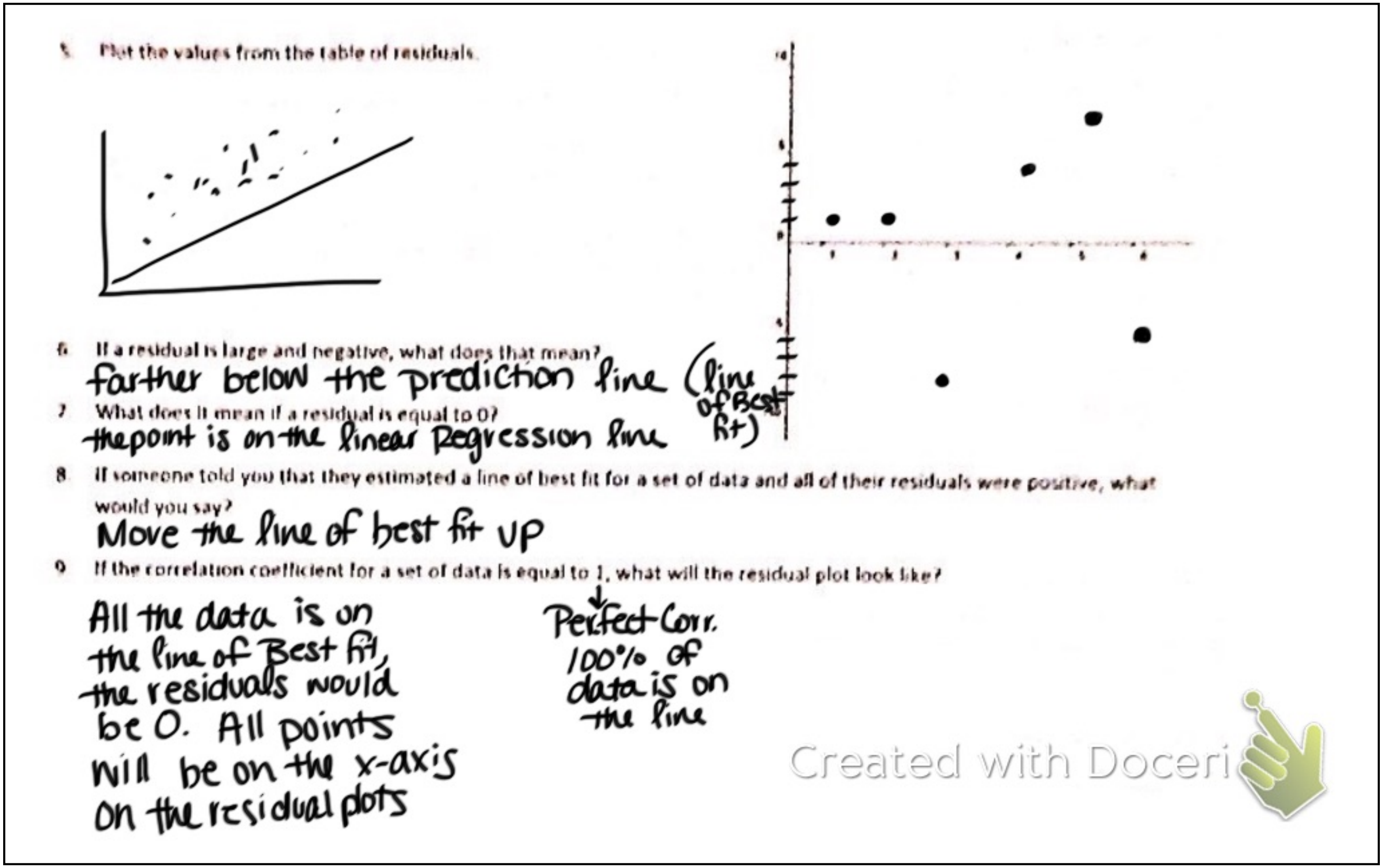
Here is the plot with the residuals drawn.

4. Use this information to complete the table below.

Plug in x → into y=3x+6

X	1	2	3	4	5	6
Observed Value	10	13	7	22	28	19
Predicted Value	9	12	15	18	21	24
Residual (observed value – predicted value)	1	1	-8	4	7	-5

data



5. Plot the values from the table of residuals.

6. If a residual is large and negative, what does that mean?

farther below the prediction line (line of Best fit)

7. What does it mean if a residual is equal to 0?

the point is on the linear Regression line

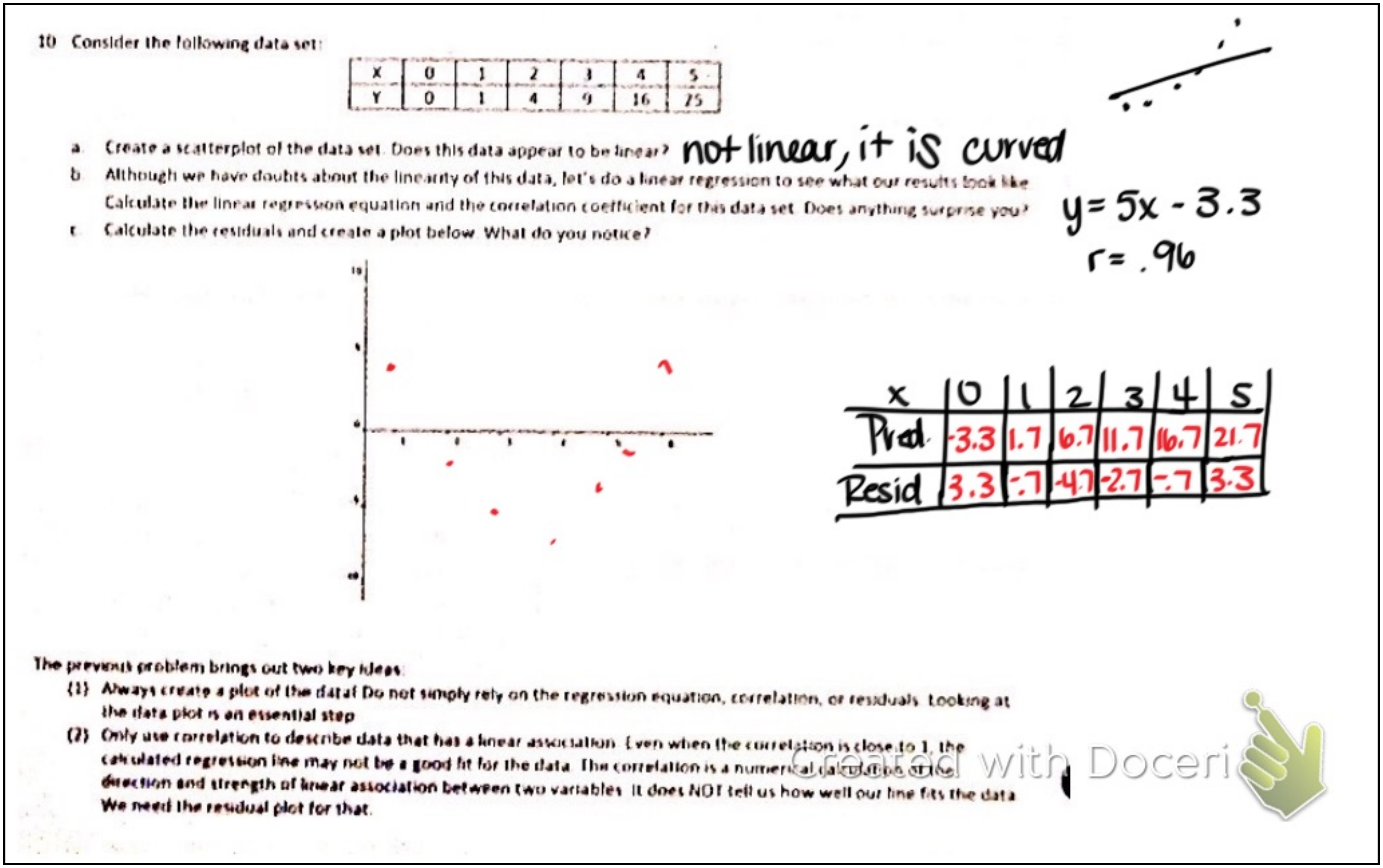
8. If someone told you that they estimated a line of best fit for a set of data and all of their residuals were positive, what would you say?

Move the line of best fit up

9. If the correlation coefficient for a set of data is equal to 1, what will the residual plot look like?

All the data is on the line of Best fit, the residuals would be 0. All points will be on the x-axis on the residual plots

Perfect Corr. 100% of data is on the line



10 Consider the following data set:

X	0	1	2	3	4	5
Y	0	1	4	9	16	25

a. Create a scatterplot of the data set. Does this data appear to be linear? not linear, it is curved

b. Although we have doubts about the linearity of this data, let's do a linear regression to see what our results look like. Calculate the linear regression equation and the correlation coefficient for this data set. Does anything surprise you?

$y = 5x - 3.3$

$r = .96$

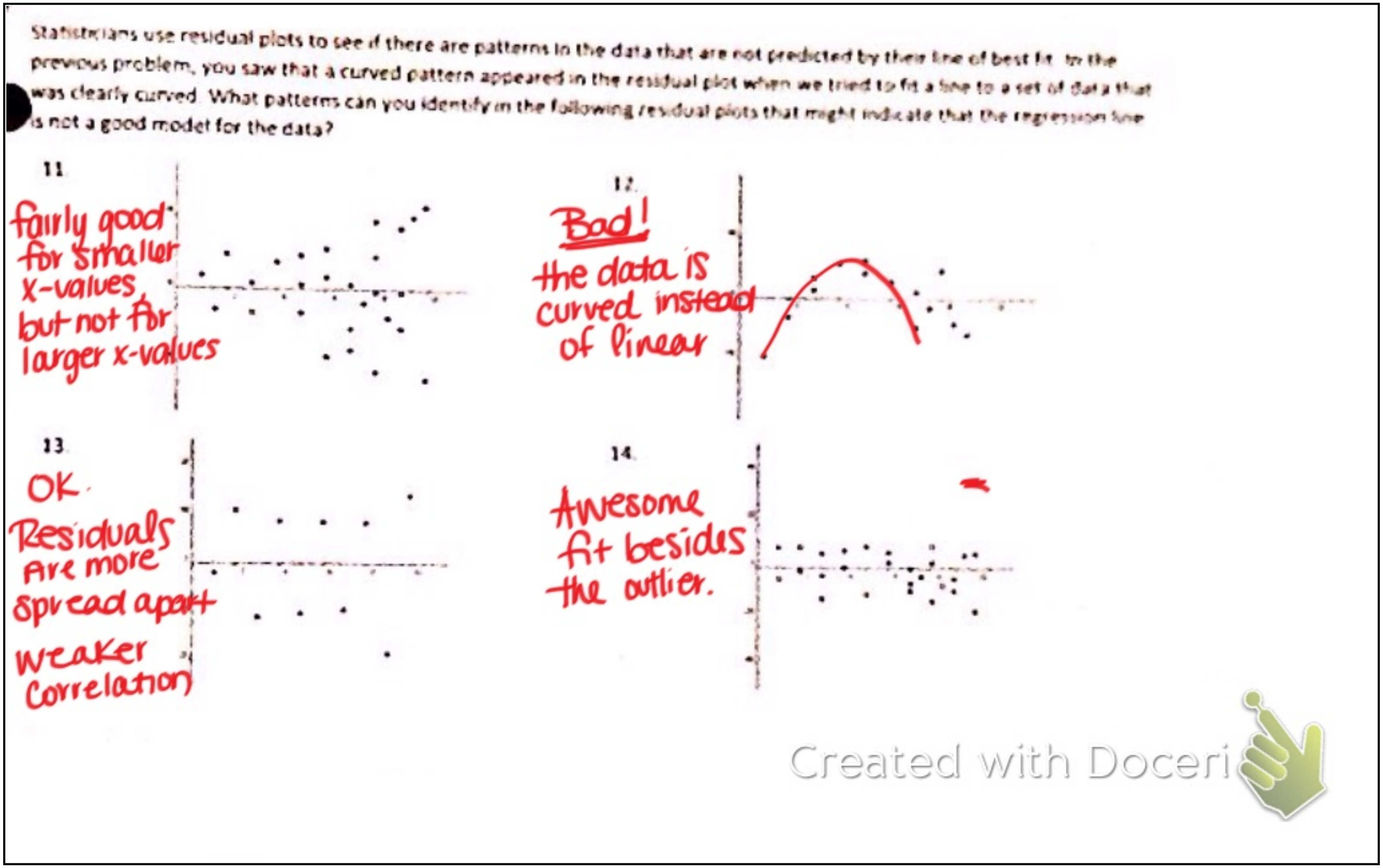
c. Calculate the residuals and create a plot below. What do you notice?

x	0	1	2	3	4	5
Pred	-3.3	1.7	6.7	11.7	16.7	21.7
Resid	3.3	-.7	-4.7	-2.7	-.7	3.3

The previous problem brings out two key ideas:

(1) Always create a plot of the data! Do not simply rely on the regression equation, correlation, or residuals. Looking at the data plot is an essential step.

(2) Only use correlation to describe data that has a linear association. Even when the correlation is close to 1, the calculated regression line may not be a good fit for the data. The correlation is a numerical calculation of the direction and strength of linear association between two variables. It does NOT tell us how well our line fits the data. We need the residual plot for that.



Statisticians use residual plots to see if there are patterns in the data that are not predicted by their line of best fit. In the previous problem, you saw that a curved pattern appeared in the residual plot when we tried to fit a line to a set of data that was clearly curved. What patterns can you identify in the following residual plots that might indicate that the regression line is not a good model for the data?

11.

fairly good for smaller x-values but not for larger x-values

12.

Bad!

the data is curved instead of linear

13.

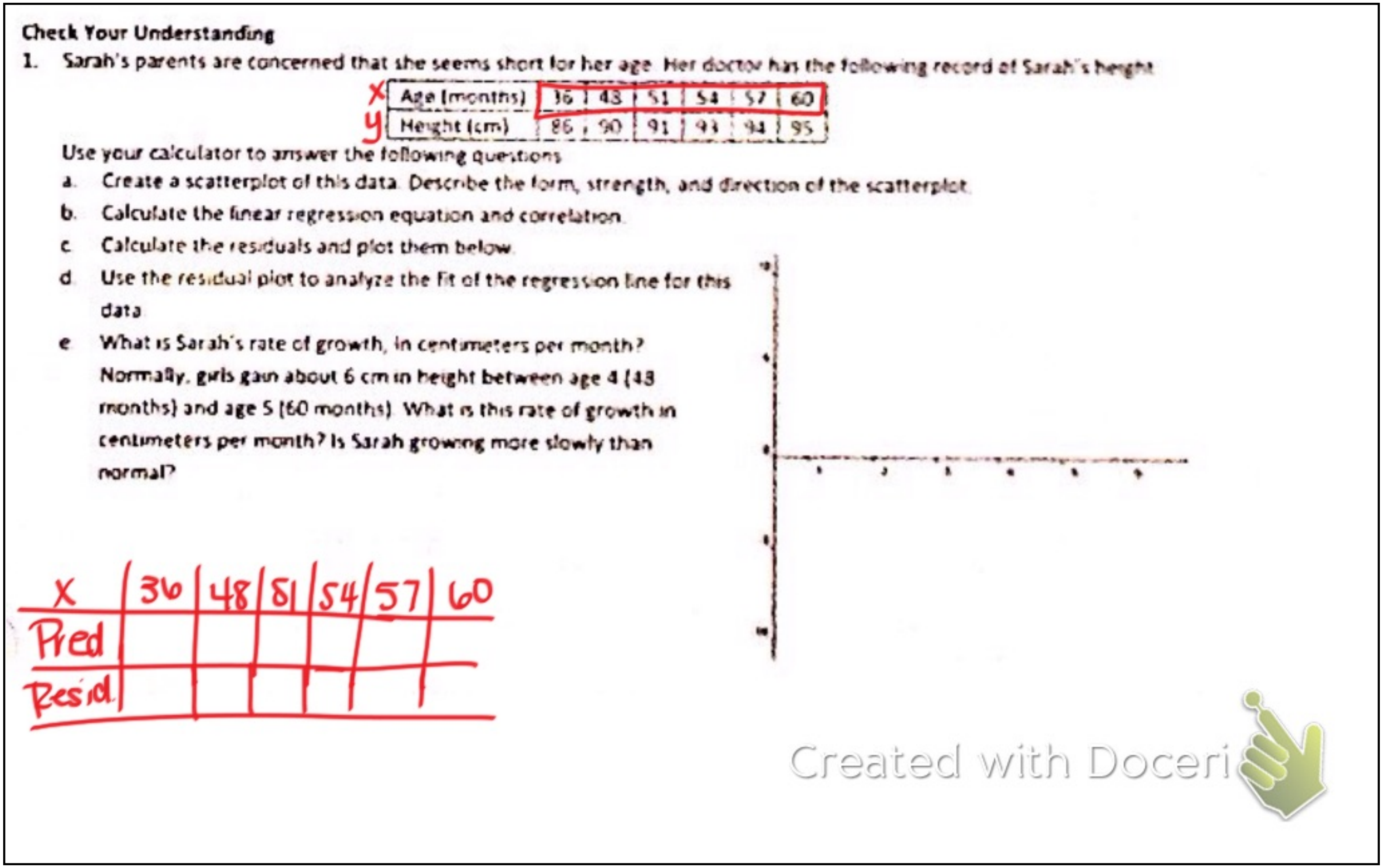
OK.

Residuals Are more spread apart

weaker Correlation

14.

Awesome fit besides the outlier.



Check Your Understanding

1. Sarah's parents are concerned that she seems short for her age. Her doctor has the following record of Sarah's height.

x	Age (months)	36	48	51	54	57	60
y	Height (cm)	86	90	91	93	94	95

Use your calculator to answer the following questions.

a. Create a scatterplot of this data. Describe the form, strength, and direction of the scatterplot.

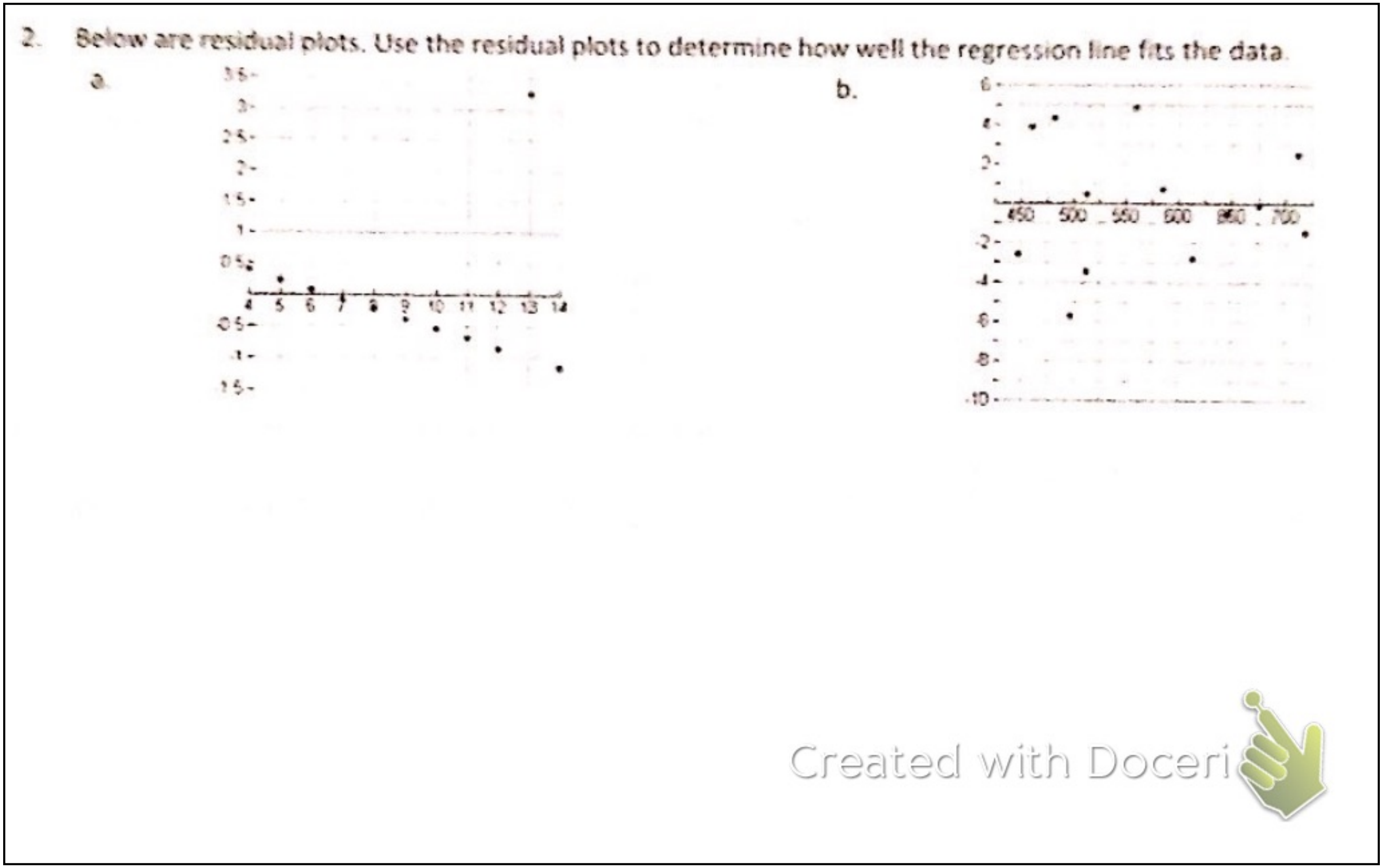
b. Calculate the linear regression equation and correlation.

c. Calculate the residuals and plot them below.

d. Use the residual plot to analyze the fit of the regression line for this data.

e. What is Sarah's rate of growth, in centimeters per month? Normally, girls gain about 6 cm in height between age 4 (48 months) and age 5 (60 months). What is this rate of growth in centimeters per month? Is Sarah growing more slowly than normal?

X	36	48	51	54	57	60
Pred						
Resid						



2. Below are residual plots. Use the residual plots to determine how well the regression line fits the data.

a.

b.