Section 10: Simple Linear Regression

Identifying Correlation Graphically

Correlation – an association or relationship between two variables

Scatter plot or **Scatter diagram** – displays the relationship between two quantitative variables.

- *x* axis – represents the *x* variable which is called the independent variable or explanatory variable
- *y* axis – represents the *y* variable which is called the dependent variable or response variable

A scatter plot is set up in a similar way to how you graph in an algebra class. The big difference is that there is not a perfect pattern so you cannot connect the points. We are simply looking for patterns in the data points themselves. The variables in a scatterplot are the same as when dealing with functions in algebra. You must always think about which variable is your *x* and which is your *y* based on which variable is dependent on the other. The difference between what is done in algebra and in this class comes from the definition of a function; each *x* will have exactly one *y*. That is not the case when dealing with data. Commonly you will give two individuals the same *x* and get different values for *y*. We must deal with the variation across individuals.

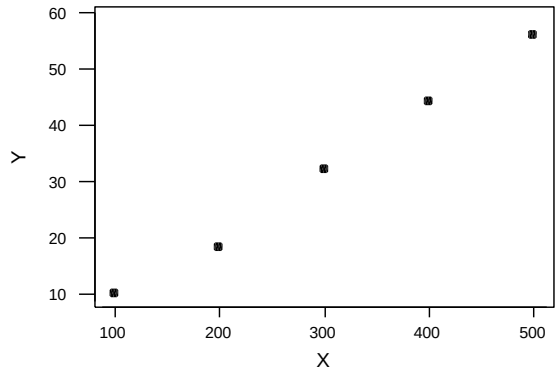
Example

Suppose you have two variables, dosage of drug and reduction in blood pressure. Which one should be your *x* and which one should by your *y*? In this case reduction in blood pressure is the *y* variable because the reduction in blood pressure would be dependent on the dosage of the drug. Therefore, the dosage of the drug is *x*, your independent variable. Below is some sample data for this situation.

x = Dosage of Drug	*y* = Reduction in Blood Pressure
100	10
200	18
300	32
400	44
500	56

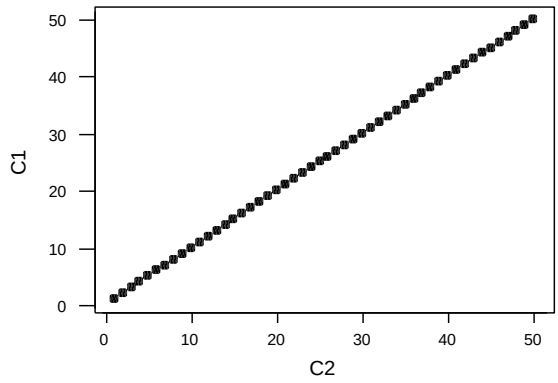
In order to graph this data, you simply plot the points on the *xy*-plane. We will not actually graph these by hand. Instead we will use software to create scatterplots. Remember that we are looking for a relationship or pattern in the data. The graph follows.

Dosage of Drug and Reduction in Blood Pressure

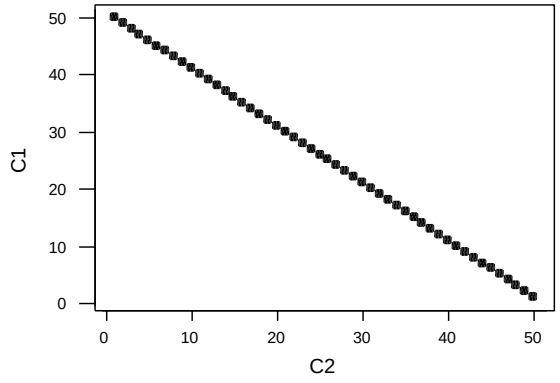


Notice in the above graph there is a pattern in the data. As the dosage of the drug increases, the reduction in blood pressure also increases. We will now look at the specific patterns you need to be able to identify.

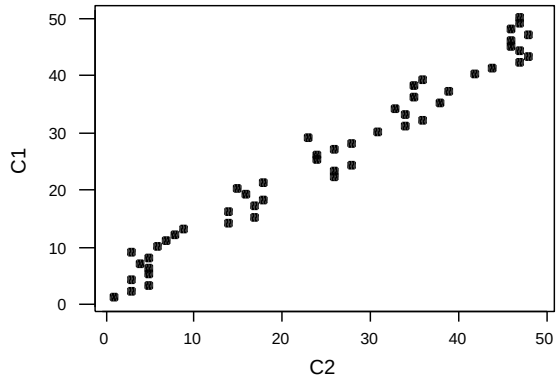
Perfect Positive Linear Correlation – We will focus on linear patterns in this class. For the graph below, all the points are exactly on a line with a positive slope. We only say perfect if all the points are exactly on a given function. Positive is used because the line has a positive slope. It is important you use these exact terms.



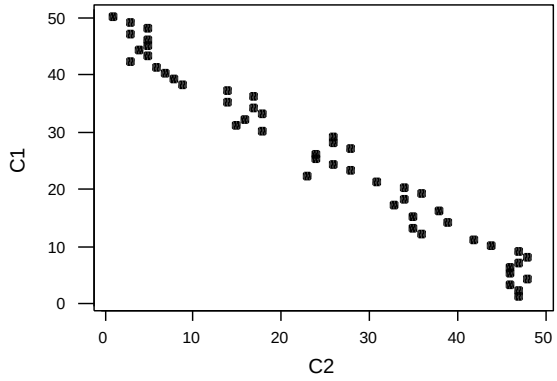
Perfect Negative Linear Correlation – The only difference between this graph and the previous one is the negative slope.



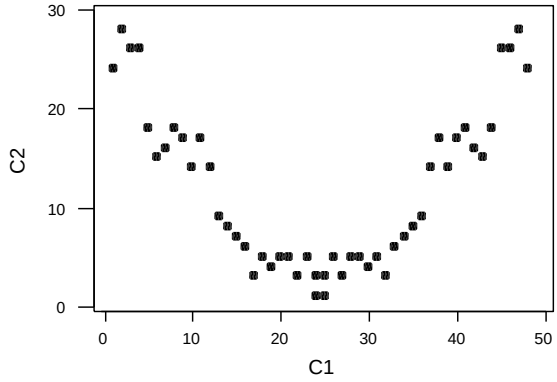
Positive Linear Correlation – The graph below is a much more reasonable situation. In practice, nothing is ever perfect. In this case there is no exact function that fits all of the points. However, if you draw a line with a positive slope, all of the points are close to the line. This is typically what we look for with this type of data.



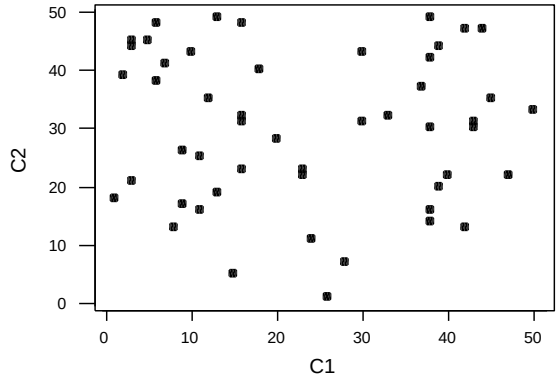
Negative Linear Correlation – Once again, the only difference between this graph and the previous one is that in this case we have a negative slope.



Non-Linear Correlation – Since our focus is on linear relationships, you do not have to identify other types of correlation. If we have a pattern that is not linear, then you should identify it as non-linear correlation. Technically the relationship below is quadratic, but you do not have to identify the specific type.



No Correlation – If there is not a pattern or relationship seen in the graph then there is no correlation as is seen in the graph below.



The Correlation Coefficient

The Pearson product-moment correlation coefficient, denoted as *r*, describes a linear relationship between two quantitative variables. It is important to notice that when looking at this value, it only indicates the linear relationship. You could have another kind of relationship present in the data. The *r* value indicates both the strength of the linear association and its direction.

The value of *r* is based on three variance components as defined below:

$$Sum\ of\ Squared\ X = SSX = \sum(x - \bar{x})^2$$
$$Sum\ of\ Squared\ Y = SSY = \sum(y - \bar{y})^2$$
$$Sum\ of\ XY = SXY = \sum(x - \bar{x})\,(y - \bar{y})$$

To calculate *r*, the following formula should be utilized:

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

Above is the progression to arrive at the computational formula which is seen last. You should use this formula when computing *r*. The algebraic steps to derive the computational formula are included below. Each variance component is manipulated first and then from those, the computational formula is derived.

$$
\begin{aligned}
SSX &= \sum (x - \bar{x})^2 \\
&= \sum (x^2 - 2\bar{x}x + \bar{x}^2) \\
&= \sum x^2 - \sum 2\bar{x}x + \sum \bar{x}^2 \\
&= \sum x^2 - 2\bar{x} \sum x + n \cdot \bar{x}^2 \\
&= \sum x^2 - 2\left(\frac{\sum x}{n}\right) \sum x + n\left(\frac{\sum x}{n}\right)^2 \\
&= \sum x^2 - 2\frac{(\sum x)^2}{n} + \frac{(\sum x)^2}{n} \\
&= \sum x^2 - \frac{(\sum x)^2}{n}
\end{aligned}
$$

$$
\begin{aligned}
SSY &= \sum (y - \bar{y})^2 \\
&= \sum (y^2 - 2\bar{y}y + \bar{y}^2) \\
&= \sum y^2 - \sum 2\bar{y}y + \sum \bar{y}^2 \\
&= \sum y^2 - 2\bar{y} \sum y + n \cdot \bar{y}^2 \\
&= \sum y^2 - 2\left(\frac{\sum y}{n}\right) \sum y + n\left(\frac{\sum y}{n}\right)^2 \\
&= \sum y^2 - 2\frac{(\sum y)^2}{n} + \frac{(\sum y)^2}{n} \\
&= \sum y^2 - \frac{(\sum y)^2}{n}
\end{aligned}
$$

$$
\begin{aligned}
SXY &= \sum (x - \bar{x})\,(y - \bar{y}) \\
&= \sum (xy - \bar{y}x - \bar{x}y + \overline{xy}) \\
&= \sum xy - \sum \bar{y}x - \sum \bar{x}y + \sum \overline{xy} \\
&= \sum xy - \bar{y} \sum x - \bar{x} \sum y + n \cdot \overline{xy} \\
&= \sum xy - \left(\frac{\sum y}{n}\right) \sum x - \left(\frac{\sum x}{n}\right) \sum y + n\left(\frac{\sum x}{n}\right)\left(\frac{\sum y}{n}\right) \\
&= \sum xy - \frac{(\sum x)(\sum y)}{n} - \frac{(\sum x)(\sum y)}{n} + \frac{(\sum x)(\sum y)}{n} \\
&= \sum xy - \frac{(\sum x)(\sum y)}{n}
\end{aligned}
$$

$$
\begin{aligned}
r &= \frac{SXY}{\sqrt{SSX \cdot SSY}} \\
&= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \cdot \sum (y-\bar{y})^2}} \\
&= \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}} \cdot \frac{n}{n} \\
&= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}
\end{aligned}
$$

Positive *r* suggests large values of *x* and *y* occur together and that small values of *x* and *y* occur together. This means that the slope of the line that best fits the points is positive. An example would be Experience and Salary. People with lower levels of experience tend to have lower salaries and people with more experience tend to have higher salaries.

Negative *r* suggests large values of one variable tend to occur with small values of the other variable. This means that the slope of the line that best fits the points is negative. An example would be Weight of a car and Gas mileage. Light cars tend to have higher gas mileage and heavier cars tend to have lower gas mileage.

So the sign of the *r* value tells us the direction of the relationship. Strength of the relationship is measured by the actual value of the number. By the term strength, we mean how close are the points to a line? The closer the points are to a line, the stronger the relationship. Because of the setup of *r*, the maximum value for *r* (in terms of absolute value) is 1. Below are some useful things to keep in mind.

$$-1 \leq r \leq 1$$

- If $r = 1$ then there is perfect positive linear correlation – all data are exactly on a line with positive slope
- If $r = -1$ then there is perfect negative linear correlation – all data are exactly on a line with negative slope
- If $r = 0$ then there is no linear relationship (keep in mind there could be another type of correlation)

The stronger the linear relationship, the larger $|r|$ (the closer to 1 this value will be). Generally, we will say there is a strong relationship if $|r| \geq .75$.

Example

Looking back to our example where we had *x* = Dosage of Drug and *y* = Reduction in Blood Pressure, what do you think the *r* value will be? Remember based on the scatterplot that the points had a strong positive linear relationship. Not perfect but pretty close meaning the *r* value should be close to 1. The calculation of the *r* value follows.

x	y	x^2	y^2	xy
100	10	10,000	100	1,000
200	18	40,000	324	3,600
300	32	90,000	1,024	9,600
400	44	160,000	1,936	17,600
500	56	250,000	3,136	28,000
$\sum x = 1500$	$\sum y = 160$	$\sum x^2 = 550{,}000$	$\sum y^2 = 6{,}520$	$\sum xy = 59{,}800$

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$
$$= \frac{5(59{,}800) - (1500)(160)}{\sqrt{(5(550{,}000) - (1500)^2)(5(6{,}520) - (160)^2)}}$$
$$= \frac{59{,}000}{\sqrt{(500{,}000)(7{,}000)}}$$
$$= 0.99728$$

Existence of correlation does not imply a cause and effect relationship

Example
There is a strong statistical correlation over the months of the year between ice cream consumption and the number of assaults in the U.S. The *r* value for this data is above .9.

Does this mean ice cream manufacturers are responsible for assaults?

No! The correlation occurs statistically because the hot temperatures of summer increases both ice cream consumption and assaults (High values occur at the same time and low values occur at the same time)

Thus, correlation does NOT imply causation. This is one of the biggest mistakes that I see in the interpretation of a correlation. You should always keep in mind that other factors besides cause and effect can create an observed correlation.

The Simple Linear Regression Model

Regression – A technique used to predict variables (typically difficult to measure variables) based on a set of other variables (typically easier to measure variables).

Linear Regression – Used to predict the value of *y* (the response variable), based on *x* (the explanatory variable) using a linear equation.

Example
Predict reaction time based on blood alcohol level. Reaction time is difficult to measure so instead we predict it with blood alcohol level which is easy to measure.

The linear regression model expresses *y* as a function of *x* plus random error.

Random error reflects variation in *y* values. Keep in mind we are going to measure *x*, so assuming we get a good measure there is no error in the *x* variable. However, when we go to use *x* to predict *y*, the prediction will not be exact. Therefore, there is error in the *y* variable. Graphically this error is represented by the vertical distance between the points and the line.

The linear regression model is:
$y = b_0 + b_1 x$
where b_0 is the y-intercept
b_1 is the slope

The above formula is the same format as what you should be used to from an algebra class. However, the way we denote the relationship is different. It is important you become familiar with this notation.

In order to use linear regression, we must first make sure the model is reasonable. The scatter plot and *r* should indicate a strong relationship. If the model is not reasonable, do not fit a line. It may still be possible to do regression with a more complicated model. However, if there is no relationship between the variables then regression cannot be used. Here, we will not worry about more complicated models, but you should understand that a simple linear model is just one of the many options available.

When using a linear regression model, we need the line that is the "best" fit for our data. Since our purpose will be to predict, we will want to pick the line that will minimize the error in the prediction. To accomplish this we will use the method of least-squares.

Method of Least-Squares – says that the sum of the squares of the vertical distances from the points to the line is minimized. Remember it is the vertical distance that represents the error.

To calculate the slope and y-intercept of the "best" fit line you can use the following formulas.

$$b_1 = \frac{SXY}{SSX} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{n\sum xy-(\sum x)(\sum y)}{n\sum x^2-(\sum x)^2}$$
$$b_0 = \bar{y} - b_1\bar{x} = \frac{\sum y - b_1\sum x}{n}$$

It is important to note that the numerator of b_1 is the same as the numerator of *r*. This should make since because these two values will always have the same sign (the denominator of both are always positive). Positive slope means *r* is positive and negative slope means *r* is negative. Also, notice the denominator of b_1 is the same as the first parentheses in the denominator of *r*. This is useful because if you calculate *r* first then you can just pull these two quantities to calculate b_1.

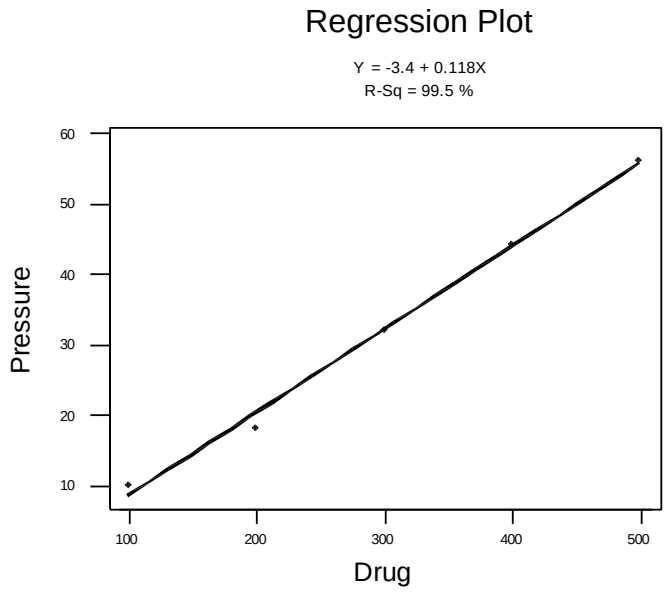
Example

Previously in this section, when looking at the correlation for dosage of drug and reduction in blood pressure we established a strong positive linear correlation (based on the scatterplot and *r*). Therefore, we can justify using linear regression in this case. Find the regression equation for the dosage of drug and reduction in blood pressure data.

$$b_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{59{,}000}{500{,}000} = 0.118$$
$$b_0 = \frac{\sum y - b_1 \sum x}{n} = \frac{160 - 0.118(1500)}{5} = -3.4$$

$$y = b_0 + b_1 x$$
$$y = -3.4 + 0.118x$$

Another way to get the regression equation is to use software. We will mainly do the calculation by hand, but if given output you should be able to identify the information needed to justify regression and predict using the model. Computer output for the dosage of drug and reduction in blood pressure data is given below.



In this output, the equation ($y = -3.4 + 0.118x$) and the R-squared value are given. If you look above the graph, you will see this information. The R-squared value is common for computer software. This value represents the percent of variation in *y* explained by the model. R-squared is always between 0% and 100%. In the linear case it is simply calculated by squaring the *r* value (R-squared $= (.99728)^2 = .995 = 99.5\%$). The higher R-squared is, the better the model. The R-squared value can be calculated for all regression models so it is commonly used to compare a linear model to other non-linear models.

You should be able to get the *r* value based on the R-squared value that is given in the output. All you have to do is take the square root of the R-squared value. The thing you have to be careful of is the direction of the relationship. Remember, if the slope of the line is positive then *r* is positive and if the slope is negative then *r* is negative. Therefore, you must look at the slope in order to decide if *r* is positive or negative.

In terms of the equation, you need to be able to use it for prediction. This is a pretty direct process as we will always be predicting *y* based on *x*. Therefore, you will plug in for *x* and solve for *y*.

For our example, predict the Reduction in Blood Pressure if 250 is the Dosage of Drug.
$y = -3.4 + 0.118x = -3.4 + 0.118(250) = 26.1$

Residual – the difference between an actual value and the fitted value (a measure of error, denoted by *e*)

$e = y - (b_0 + b_1 x)$

Example
Find the residual value for the point (400, 44) in the dosage of drug and reduction in blood pressure data.

$e = y - (b_0 + b_1 x) = 44 - \big(-3.4 + 0.118(400)\big) = 44 - 43.8 = 0.2$

This tells us that the point (400, 44) is 0.2 units above the "best" fit line.

Cautions with regression

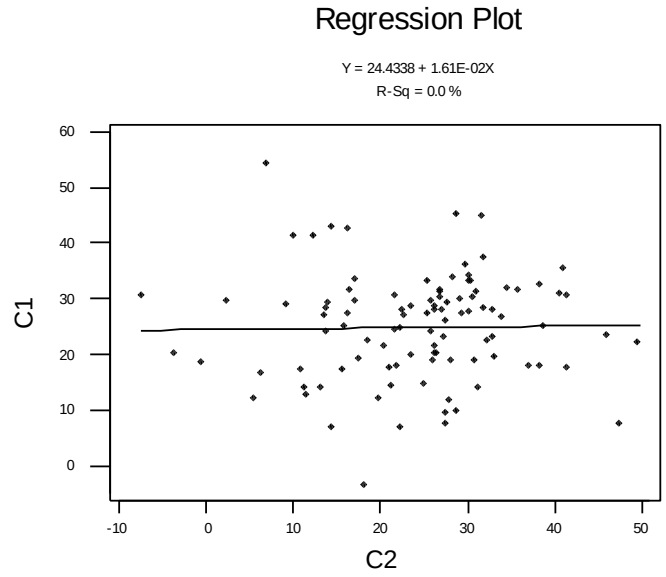
There are two common mistakes with regression. You must be aware of the problems with extrapolation and extreme values.

Interpolation – predicting Y values for X values that are within the range of the scatter plot (this is what regression should be used for)

Extrapolation – predicting Y values for X values beyond the range of the observations (this should not be done with a basic regression model, it is a complex problem)

If our *x* variable ranges from 100 to 500 as it does in the dosage of drug and reduction in blood pressure example then it is reasonable to predict within that range (interpolation). However, if you try and predict for an *x* of 1000 (extrapolation) then you have no data indicating that this relationship holds at that value. It is quite possible that the relationship changes beyond the range of the data. There is no way to know this without collecting data consistent with the *x* values you want to predict. Regression should not be used to extrapolate.

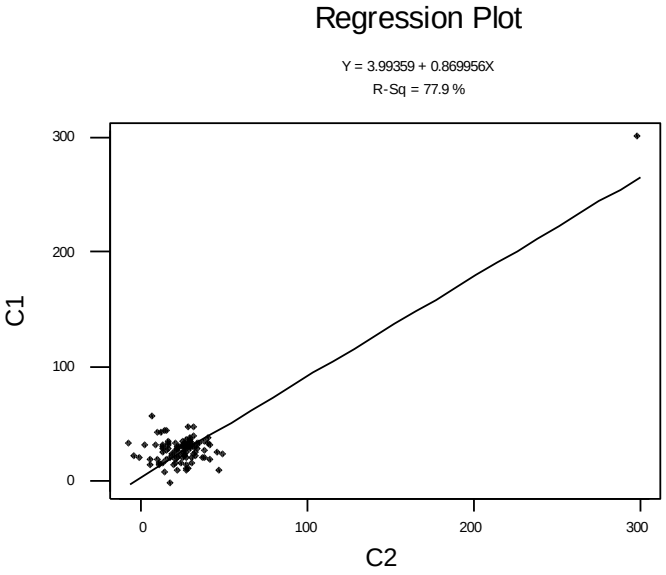
The least-squares line and the *r* value can be affected greatly by extreme data points. In order to illustrate this we will look at some computer output.



Calculate the *r* value for the above data.

$r = \sqrt{0} = 0$

With an *r* of 0, we know that there is no linear relationship between *x* and *y*.



Calculate *r* for the above data.

$r = \sqrt{.779} = .883$

With an *r* of .883, which is bigger than the criteria of .75 it seems like we have a strong relationship. With further investigation via the scatterplot, you will see that all of the data is in the bottom left of the graph except one data point which is extreme. What I actually did was take the data from the previous graph with an *r* value of 0 and add one extreme value. Notice the extreme value makes the other data points appear close together. They also appear numerically close since the one value is so extreme. Therefore, the *r* value is high because the points are close to the line. In this case linear regression is not justified. If you have an extreme value in a plot like in this case, you should remove the extreme value and see if the relationship still exists. In this case it does not, so linear regression will not work for this data.