Correlation and the intuition behind simple linear regression

So far you've learned that simple linear regression is a technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y. You've also learned about ordinary least squares estimation (OLS), which is a common way to determine the coefficients of the regression line—the line of "best fit" through the data. In this reading, you'll explore the meaning of correlation; learn about *r*, or the "correlation coefficient;" and discover how to determine the regression equation. This knowledge will help you better understand relationships between variables, and thus how linear regression works.

Correlation

Correlation is a measurement of the way two variables move together. If there is a strong correlation between the variables, then knowing one will be very helpful to predict the other. However, if there is a weak correlation between two variables, then knowing the value of one will not tell you much about the value of the other. In the context of linear regression, correlation refers to *linear* correlation: as one

variable changes, so does the other at a constant rate.

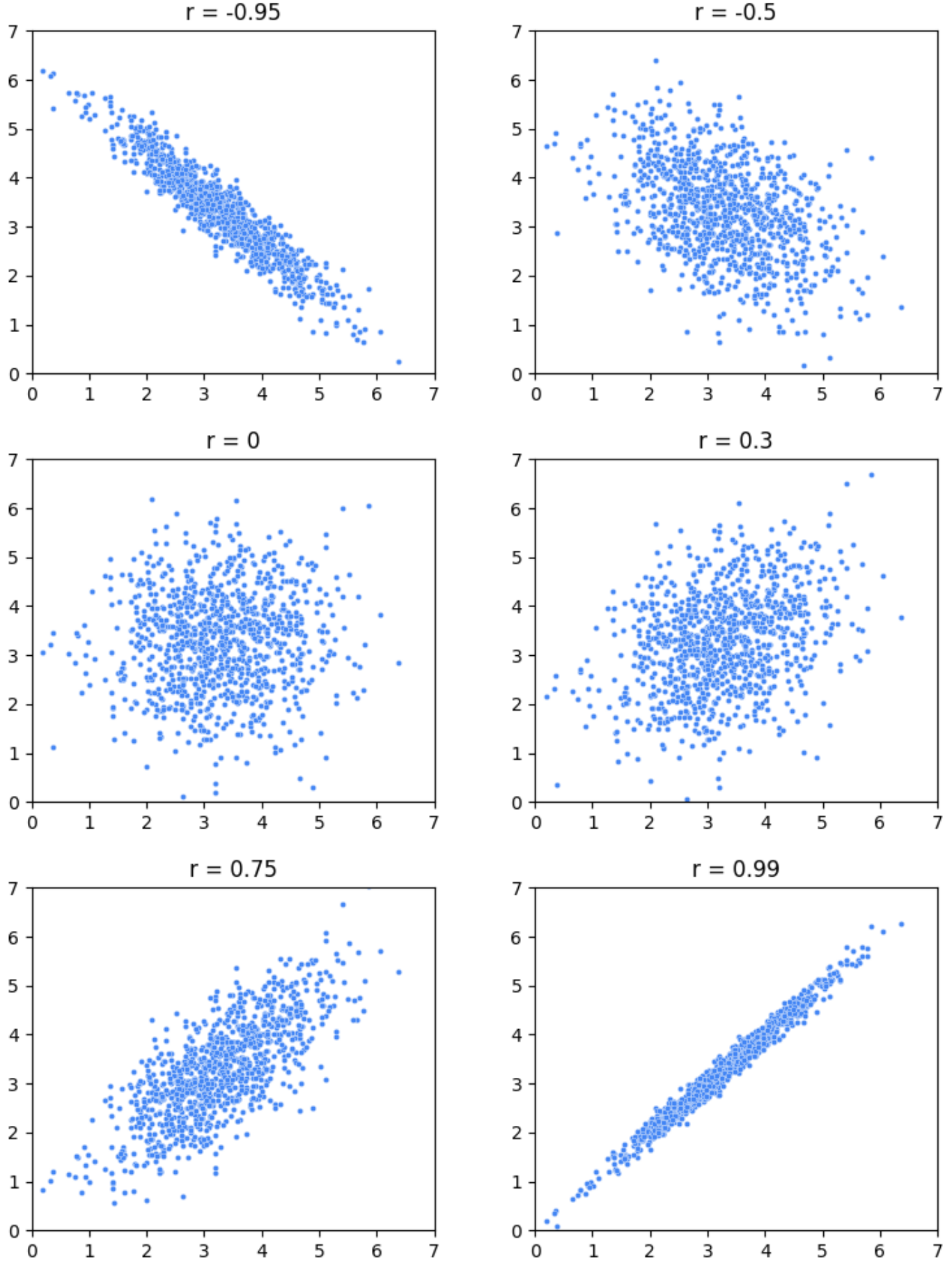
In the statistics course, you learned that a continuous variable can be summarized using some basic numbers. Two of these summary statistics are:

- **Average:** A measurement of central tendency (mean, median, or mode)
- **Standard deviation:** A measurement of spread

When two variables are summarized together, there is another relevant statistic called ***r***, **Pearson's correlation coefficient** (named after the person who helped develop it), or simply the linear **correlation coefficient.** The correlation coefficient quantifies the strength of the linear relationship between two variables. It always falls in the range of [-1, 1]. When *r* is negative, there is a negative correlation between the variables: as one increases, the other decreases. When *r* is positive, there is a positive correlation between the variables: as one increases, so too does the other. When *r* = 0, there is no *linear* correlation between the variables. Note that there are cases where one variable might be precisely determined by another—like $y=x^2$ or y=sin(x)—but the value of the *linear* correlation between X and Y would nonetheless be low or zero because their relationship is non-linear.

The following figure depicts scatterplots of bivariate (bi =

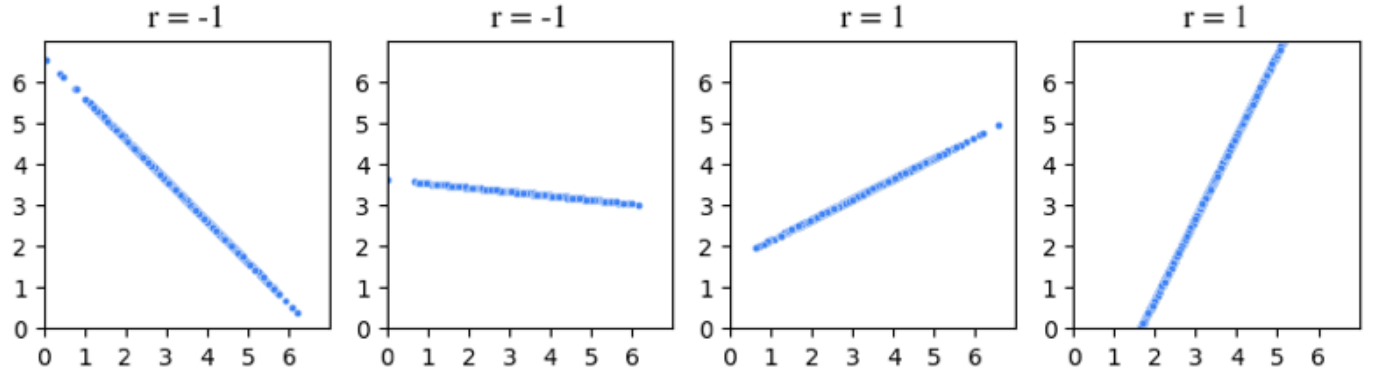
“two”, variate = “variables”) data where each variable has the same mean and standard deviation and only the correlation coefficient varies.



Notice that the closer to -1 or 1 *r* is, the more linear the data

appears. When *r* is exactly 1 or exactly -1, then the variables are perfectly correlated, and their graph is a line. When *r* is zero, there is no correlation between the variables, and, in this example, the data appears as a shapeless cloud of points.

However, *r* only tells you the strength of the linear correlation between the variables; it does not tell you anything about the magnitude of the slope of the relationship between the variables aside from its sign. For example, variables with *r*=1 wouldn't tell you if increasing X by one would lead to Y increasing by 10, 100, 0.1, or something else. It would only tell you that you can be sure that it *would* increase. This fact is illustrated in the following figure, where even though the slopes of the lines are all different, *r* is only either -1 or 1. If the line is perfectly horizontal or perfectly vertical, then *r* is undefined. (If you're wondering why, refer to the equation below. One of the terms in the denominator would equal zero, which would make the whole denominator equal zero, which would result in an undefined solution.)



Calculate *r*

The formula for *r* is:

$$r = (SD\ X)(SD\ Y) covariance(X,Y)$$

where:

$$covariance(X,Y) = ni = 1 \sum n(xi - \bar{x})(yi - \bar{y})$$

Note: The formulas for *r* and covariance given here represent those used for entire populations. For samples, the denominator of the covariance formula is *n - 1* and, similarly, the standard deviations in the formula for *r* are calculated using *n - 1* instead of *n*. For simplicity, this reading will use the population formulas in its demonstrations.

An easier way of thinking about this calculation is: the numerator—the covariance—represents the extent to which X and Y vary together from their respective means. When this value is positive, it suggests that high values of X tend to be associated with high values of Y, indicating a positive correlation. Conversely, if the value is negative, it suggests that high values of X tend to be associated with low values of Y and vice versa, indicating a negative correlation.

The denominator—the product of the standard deviations—standardizes the units of the numerator. It adjusts for the inherent variability of the individual variables. This makes *r* a

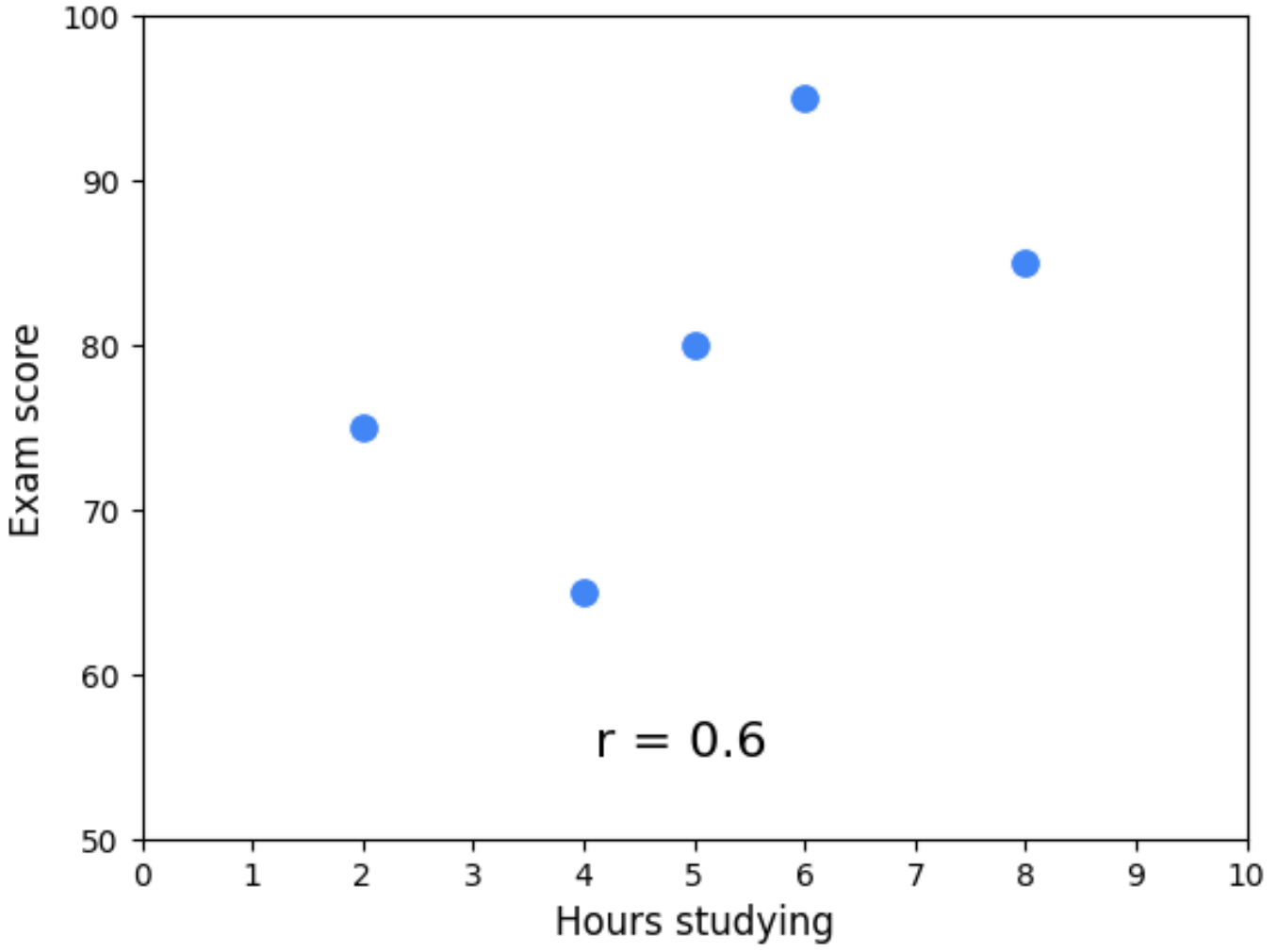
statistic without a unit. It is a pure number, without dimension.

An equivalent way to calculate *r* is to convert each data point in each variable to standard units (subtract the mean, divide by the standard deviation), then take the average of the products.

Here's an example. Suppose five students took an exam and you recorded how many hours they spent studying and also their grade. The following table breaks out the calculation of *r*.

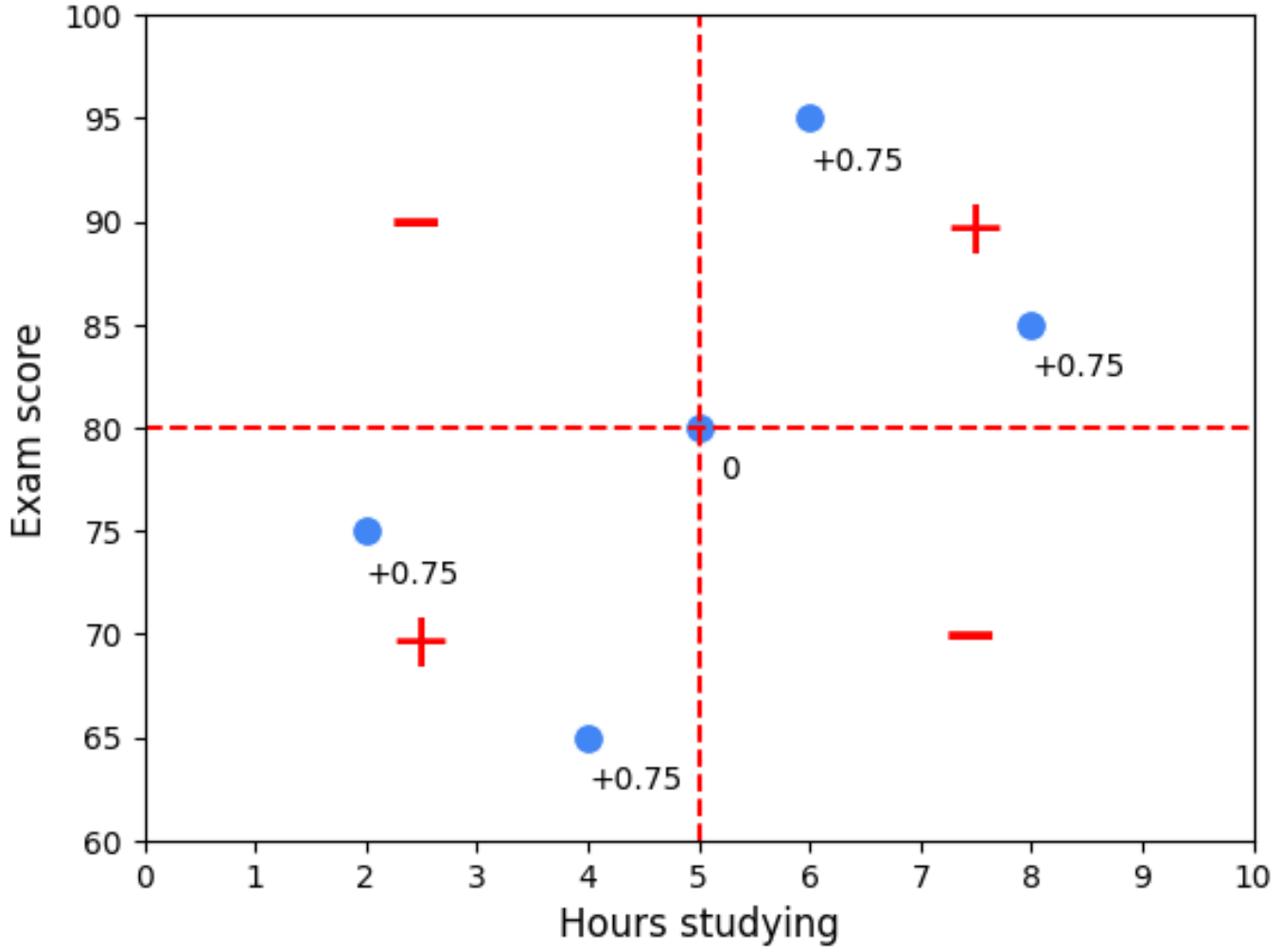
Hours studying (X)	**Exam grade (Y)**	**X in standard units**	**Y in standard units**	**Product of standard units**
2	75	-1.5	-0.5	0.75
4	65	-0.5	-1.5	0.75
5	80	0	0	0
6	95	0.5	1.5	0.75
8	85	1.5	0.5	0.75
mean X = 5 **SD X = 2**	**mean Y = 80** **SD Y = 10**			**mean of products (r) = 0.6**

The correlation coefficient is 0.6. Here is a graph of this data:



Notice that the cloud of points slopes upwards. This corresponds with *r* being positive. The correlation coefficient works as an indicator of association because it uses the product of each variable's deviation from its mean. When the product is positive, it means *both* the X and the Y values are either below their respective means (negative standard units) or above their respective means (positive standard units). They vary together. However, when this product is negative, it means one of the values is above its mean and the other is below it. They vary in opposing directions relative to their respective means.

The following figure illustrates this idea. The figure is divided into quadrants. The vertical line represents the mean X value and the horizontal line represents the mean Y value. Each point is labeled with the product of its standardized scores (refer to the table above). The average of these scores is *r*. When *r* is positive, more points will tend to be in the positive quadrants, and vice versa.

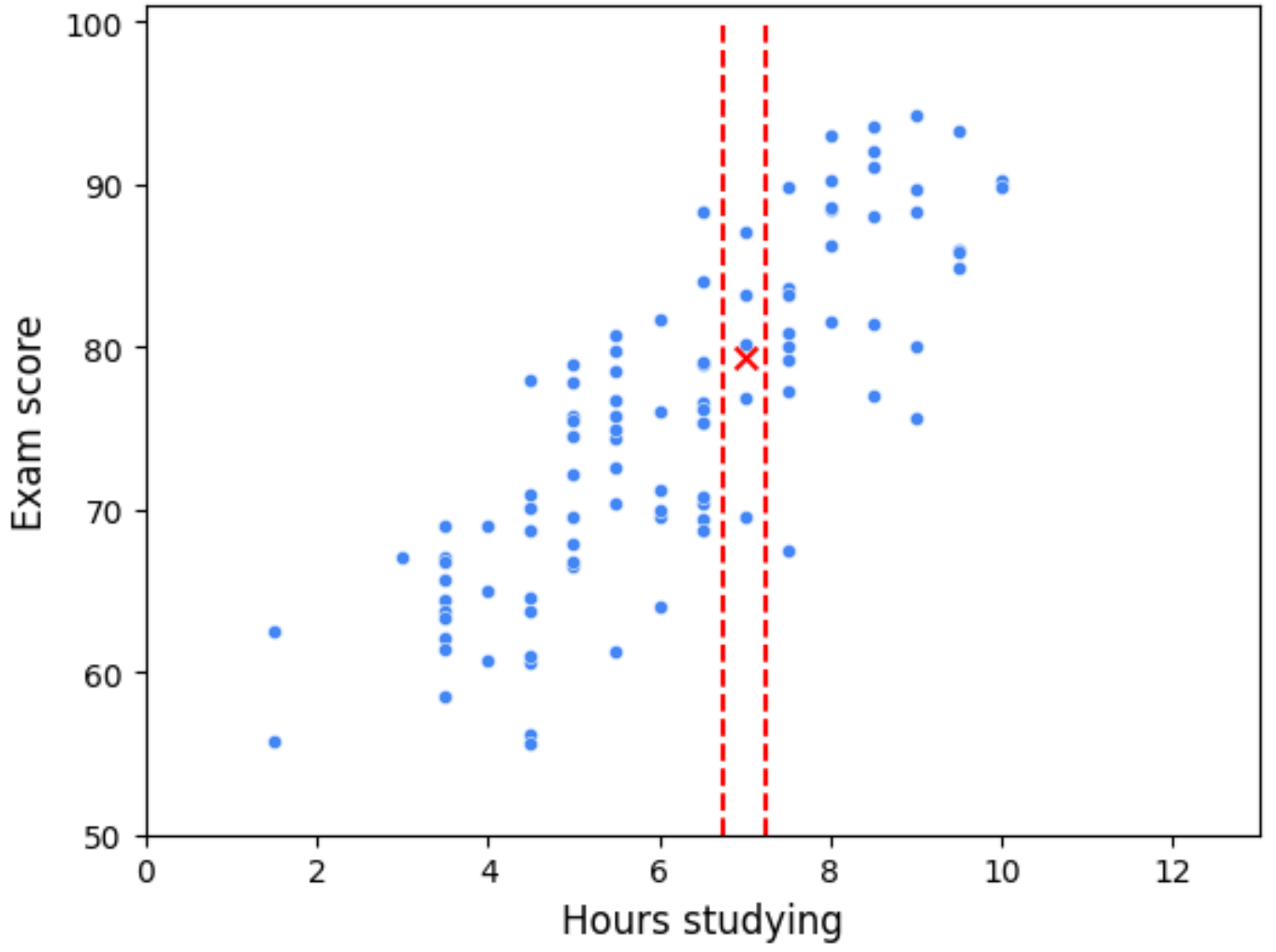


Regression

In the absence of any other information, if you had to guess a randomly selected student's exam score, the best way for you to minimize your error would be to guess the average of

all the students' scores. But what if you also knew how many hours that student studied? Now, your best guess might be the average score of only the students who studied for that many hours.

Here is an example using a sample of 100 students with study times rounded to the nearest half hour. Suppose you were told a student studied for seven hours. To guess their exam score, one way to minimize error is to guess the average of only the students who studied for seven hours.



In this scatterplot, all of the students who studied for seven hours fall between the two vertical lines. Their mean exam

score is represented by an X. Linear regression expands on this concept. A regression line represents the estimated average value of Y for every value of X, given the assumptions and limitations of a linear model. In other words, the actual average Y values for each X might not lie exactly on the regression line if the relationship between X and Y is not perfectly linear or if there are other factors influencing Y that are not included in the model. The regression line attempts to balance out these influences to find a straight-line relationship that best fits the data as a whole. It's an estimation of the central tendency of Y, given X.

The regression equation

Now that you know about *r* and you better understand the concept of regression, you're ready to put everything together to find the line of best fit through the data. The formula for this line is known as the regression equation. There are two keys to this step.

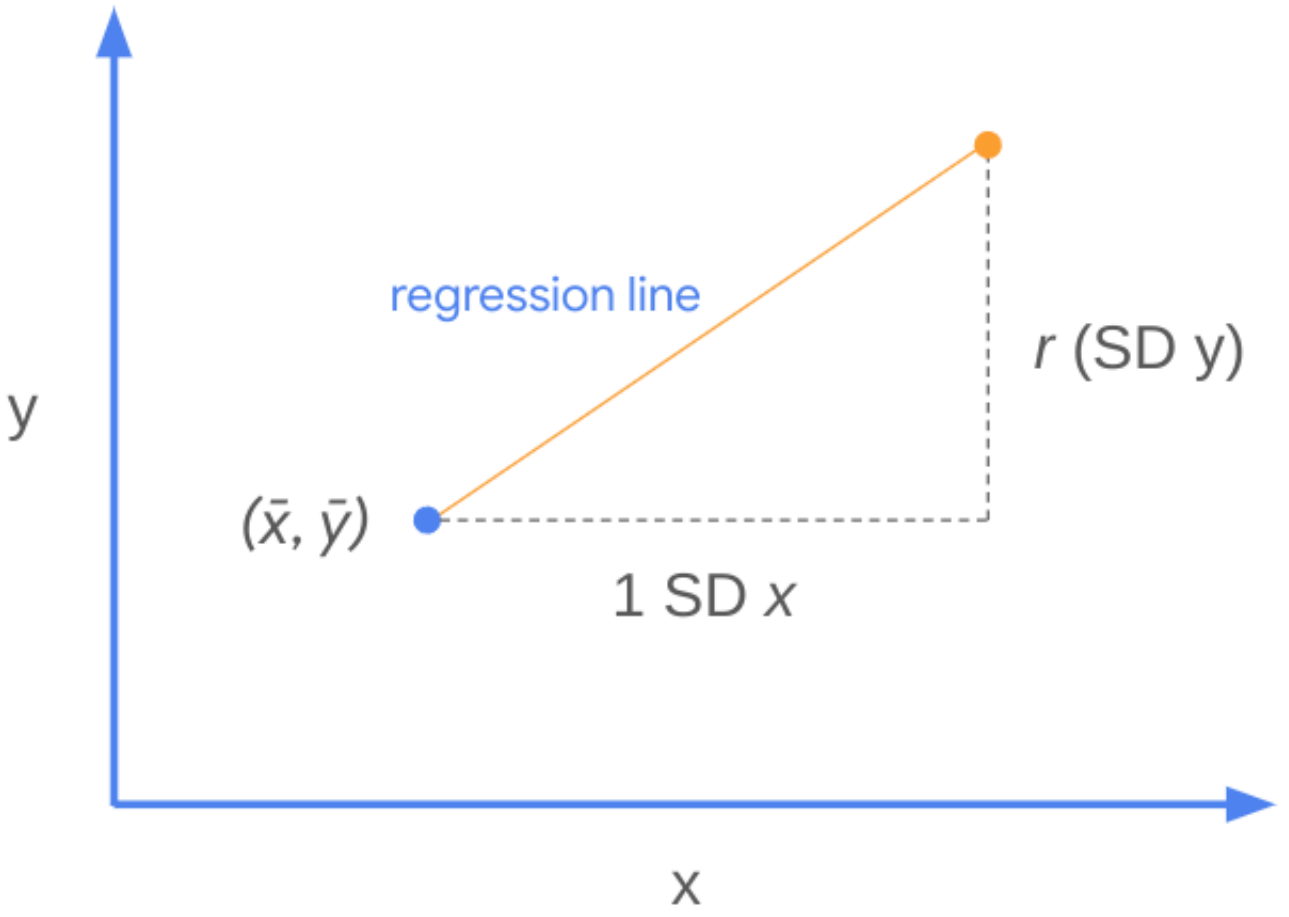
The first is:

- The mean value of X and the mean value of Y (i.e., point $(\bar{x}, \bar{y})$) will always fall on the regression line.

The second is to understand what *r* means:

- For each increase of one standard deviation in X, there is an expected increase of *r* standard deviations in Y, on average over X.

The following figure illustrates how these concepts work together to determine the regression line.



In other words, the slope of the regression line is:

$m = SD\ x r(SD\ y)$

This is *m* in the formula for a line: $y = mx + b$. The intercept, represented by *b,* is therefore: $b = y - mx$. Because you know that point $(\bar{x}, \bar{y})$ is always on the regression line, you can plug

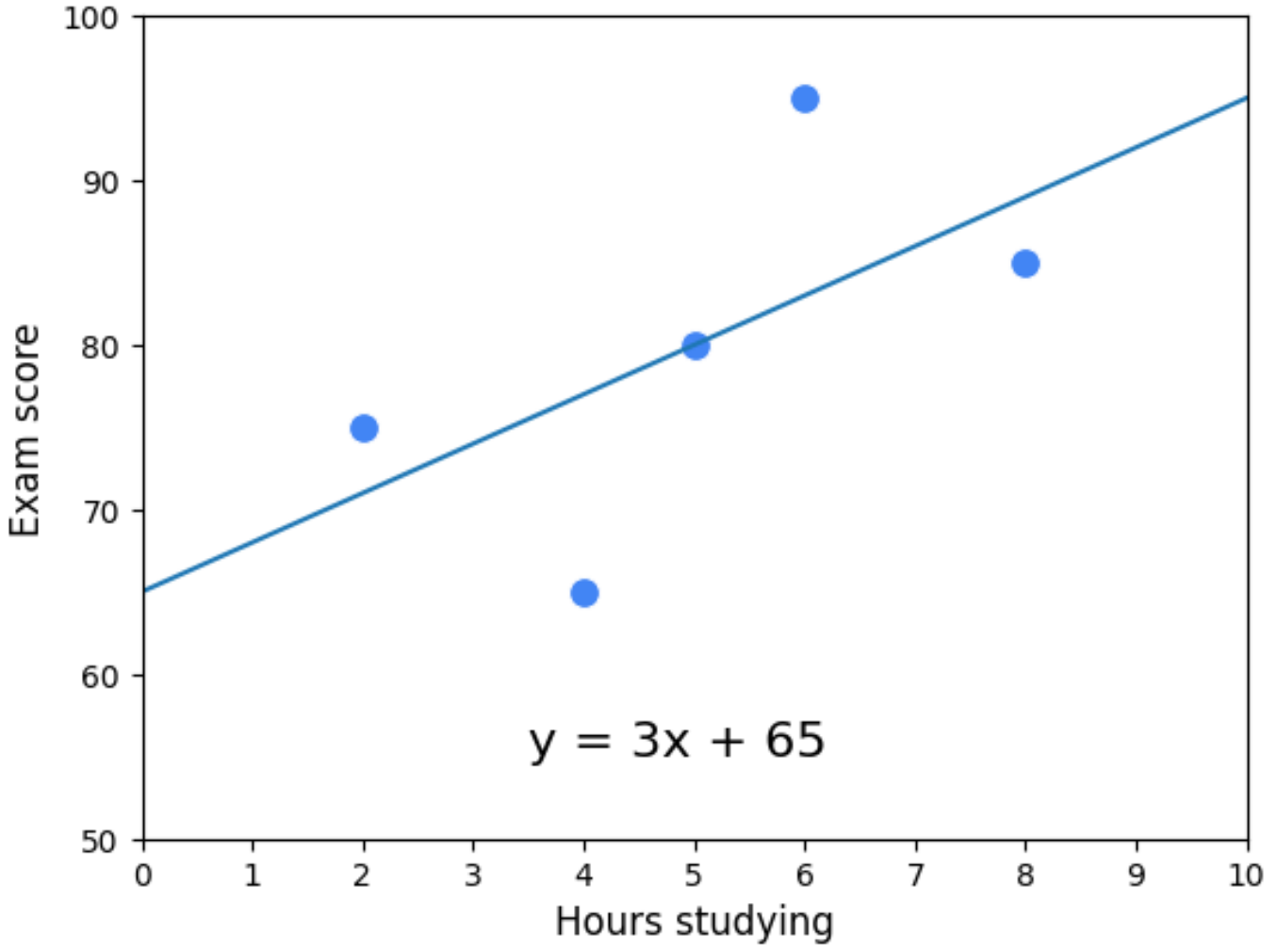
in the x and y values from this point to calculate the intercept. Here's an example using the original sample of five students.

	Hours studying (X)	**Exam grade (Y)**
mean:	5	80
SD:	2	10
r:	**0.6**	

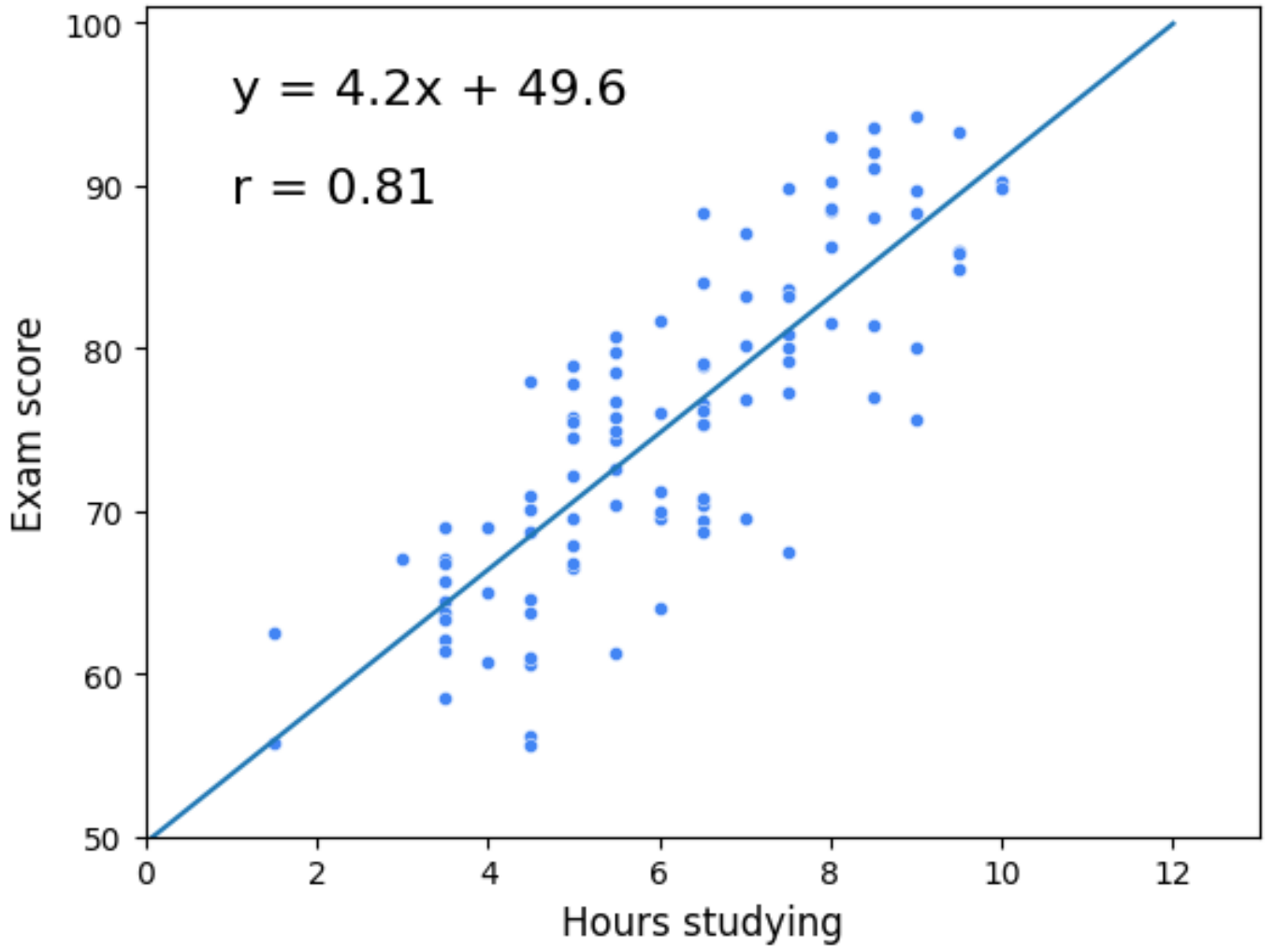
Broken into steps:

1. Calculate slope: $m = SD\ x r(SD\ y) = 20.6(10) = 3$.
2. Calculate the intercept: Substitute $\bar{x}$, $\bar{y}$, and m into the equation $y = mx + b$: $80 = 3(5) + b \rightarrow b = 65$.
3. Generalize to get the regression equation: $y = 3x + 65$.

Here is the regression line overlaid onto the data:



This is referred to as "the regression of Y on X." Here is the regression line for all 100 students:



Key Takeaways

Linear regression is one of the most important tools that data professionals use to analyze data. Understanding the fundamental building blocks of simple linear regression will help you as you continue learning about more complex methods of regression analysis. Here are some key points to keep in mind:

- Correlation is a measurement of the way two variables move together.
- *r (a.k.a.* Pearson's correlation coefficient, a.k.a.

correlation coefficient) quantifies the strength of the linear relationship between two variables.

    - It always falls in the range of [-1, 1].
    - Variables that tend to vary together from their means are positively correlated. Conversely, variables that tend to vary in opposite ways to their respective means are negatively correlated.

- The regression line estimates the average y value for each *x* value. It minimizes the error when estimating *y*, given x.
- The slope of the regression line is *SD xr*(*SD y*).
- The point ($\bar{x}$, $\bar{y}$) is always on the regression line.