MATH 201 Week 5 - Linear Regression and Correlation

In most research, it is very common to attempt to determine the extent to which one variable (the independent variable) has an effect on another variable (the dependent variable). While the actual cause and effect relationship is not determined, we can use these techniques to determine the magnitude of the relationship between the variables.

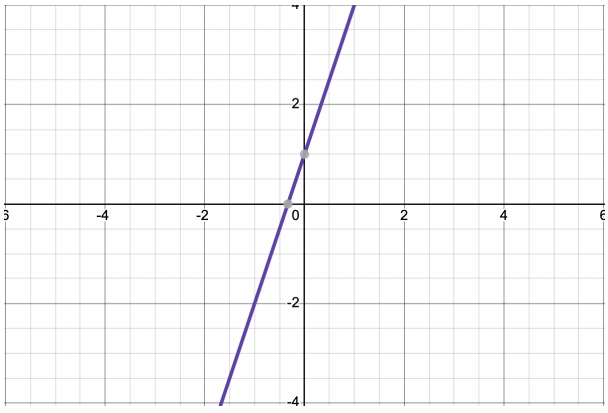
Linear Equations

To begin this exploration, we start with a review of the equation of a line. In the following equation, x is the independent variable, y is the dependent variable, a is the y-intercept value, and b is the slope of the line.

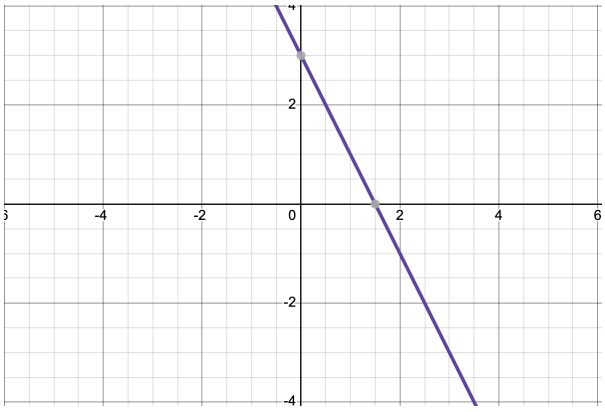
$$y = a + bx$$

This is the basic form of a linear equation representing the mathematical relationship between x and y. A linear equation is just the equation of a line. Extracting a linear equation from the comparison of sample data allows the researcher to build a model that can be used to determine the nature of the relationship between the variables and potentially be used to predict a future condition of the system. The following are examples of linear equations:

y = 3x + 1



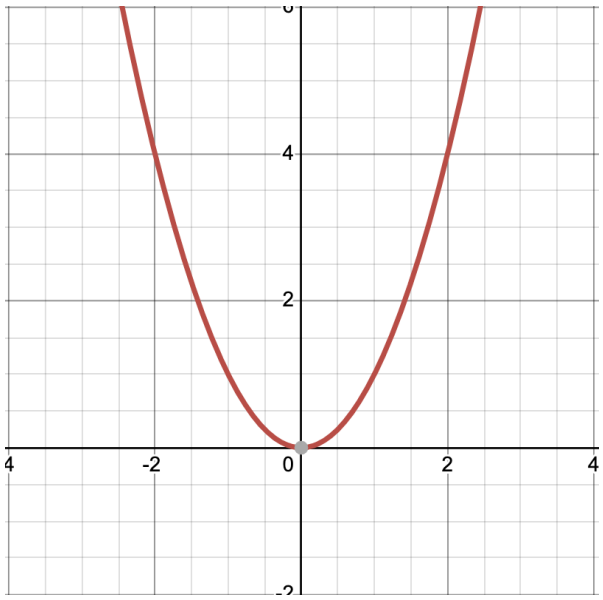
y = -2x + 3



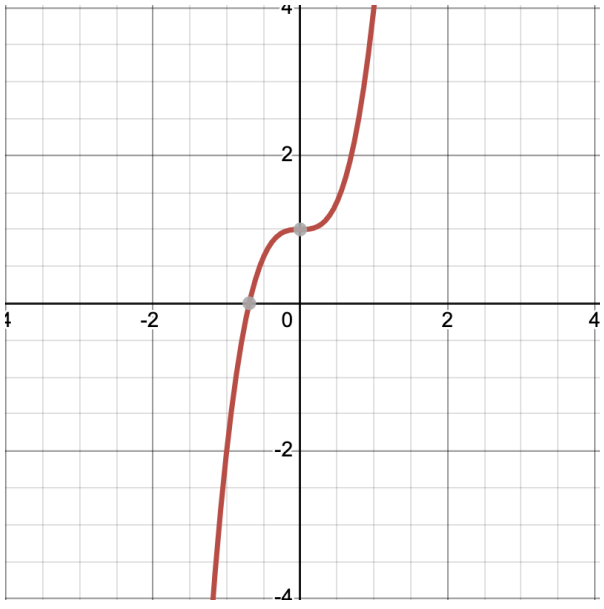
A powerful online tool for graphing linear equations, which is very helpful in visualizing these equations, can be accessed at https://www.desmos.com/calculator.

It is important to mention (again) that linear equations are equations of lines, not curves. The following examples are not linear equations because they are not represented by straight lines.

$y = x^2$



$y = 3x^3 + 1$



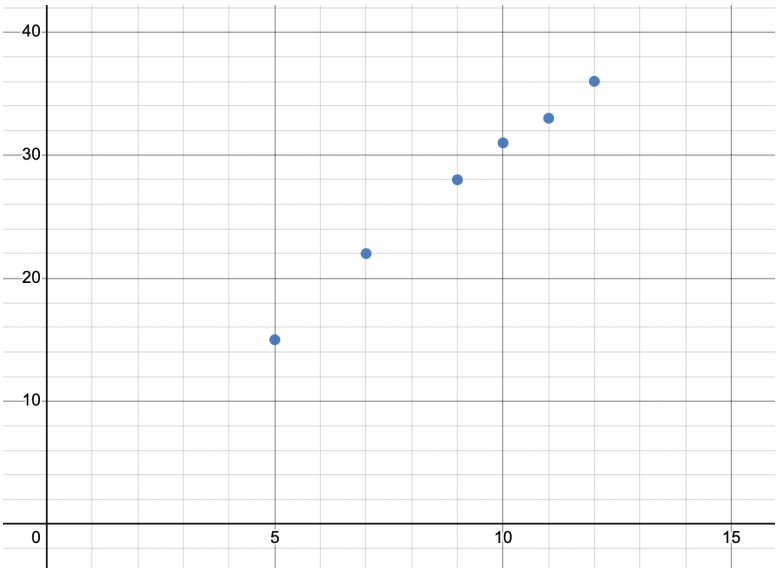
Linear Regression

Linear regression is the process of modeling the relationship between two variables by drawing a best-fit line that represents the trend revealed by a scatter plot of the data. To illustrate this, let's take a look at an example.

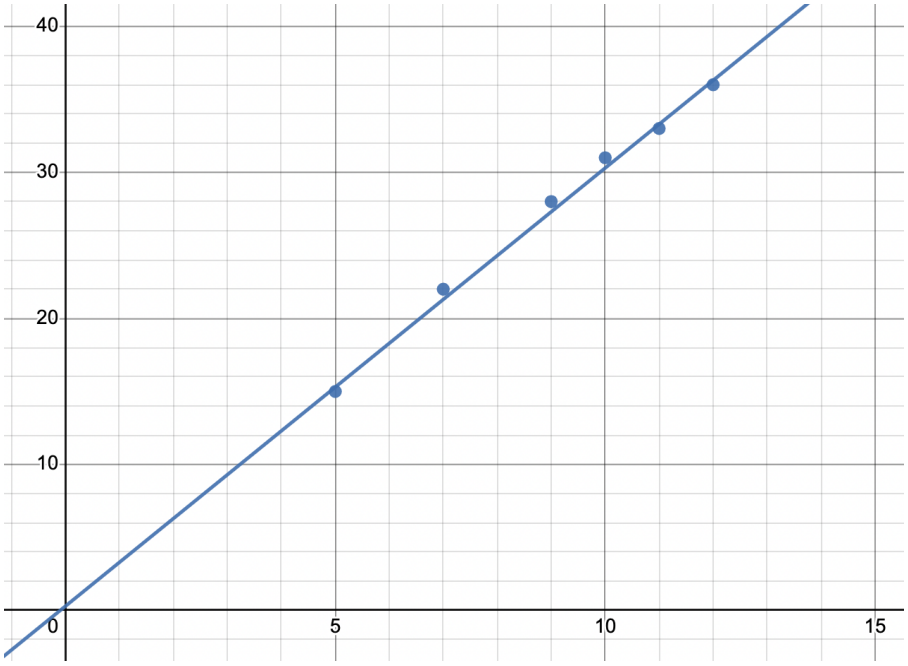
Amelia plays basketball for her high school. She wants to improve her game, so she decides to see if the number of hours that she practices (independent variable) has an impact on the number of points she scores in a game (dependent variable). The following table represents the data that Amelia has collected:

X hours practiced	Y points scored in a game
5	15
7	22
9	28
10	31
11	33
12	36

Graphing this information we obtain the following scatter plot:



The scatter plot shows that there is definitely a positive relationship between the number of hours of practice (x) and the number of points scored (y). Now, we can develop a linear equation to represent this relationship by drawing a best fit line through the points like so:



Estimating or Predicting

The linear equation for this particular pattern is $y = 3x + 0.3$. We can use this linear equation to make an estimate (or prediction) regarding how many points *(y)* Amelia would score if she practiced *(x)* hours. For example, if she practiced 15 hours she would potentially earn $y = 3(15) + 0.3 = 45.3$ or approximately 45 points in a game.

When it comes to linear regression, it is worth noting that this linear equation is not an absolute representation of the reality of this situation. It is only a model that can be used to illustrate the potential relationship between variables in order to make an informed decision.

Spreadsheets and statistical calculators are very efficient at computing the best fit lines for scatter plots. This is done by applying a least-squares regression line to obtain the best fit line. This approach attempts to limit the error (or residual) difference between points in the scatter plot and points on the best-fit line.

Outliers

It is inevitable that a sample data set will contain some outliers (due to many different factors). In the case of linear regression, outliers are data points that are far from the least squares best-fit line. As such, they are considered to have large errors (or residuals) in their vertical distance from the regression line. Dealing with outliers must be carefully considered. While it is true that some outliers might just be anomalies produced by the sampling methodology, it is also quite possible that they hold some kind of valuable information that warrants closer inspection.

Outliers can be quickly identified in scatter plots, but, as a general rule, any data point more than two standard deviations above or below the best-fit line could be flagged as an outlier. Outlines have a large effect on the slope of the best-fit line.

The Slope of the Line

Recall that the generic equation for a line is as follows: $y = a + bx$. In this equation, b is the slope of the line. The slope is basically the rate at which the value of y changes relative to the value of x. In our example above, we found the equation of the line to be $y = 3x + 0.3$. The slope of this line is 3 and we would interpret it to mean that Amelia improves her score by 3 points for every hour she practices.

Let's say we have the following linear equation that represents the relationship between number of species diversity (y) and number of acres (x) in a community: $y = -0.4x + 3$. For this equation, the slope is interpreted as for every acre lost, there is a 0.4 decline in species diversity in the community.

The Correlation Coefficient r

The correlation coefficient r (also known as Pearson's r) is a number that tells us the strength of the relationship between the independent variable x and dependent variable y. The formula for the correlation coefficient is as follows:

$$r = \frac{n\Sigma(xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

Where n = number of data points. The good news is that all spreadsheets and statistical calculators can quickly determine the value of the correlation coefficient r for you.

The value of r is always between -1 and +1 ($-1 \leq r \leq 1$). Values of r close to -1 or +1 are very strong relationships. When r = +1, there is a strong positive correlation between the variables. When r = -1 there is a strong negative correlation between the variables. It is, however, important to note that real world data seldom results in a perfectly straight line, therefore r values close to 1 and -1 are interpreted as being strong correlations. The closer the value of r is to zero, the less likely there is a correlation between the variables.

Correlation Coefficient r	Description
+1	Strong Positive Correlation
+0.5	Weak Positive Correlation
0	No Correlation
-0.5	Weak Negative Correlation
-1	Strong Negative Correlation

The graphics below demonstrate several scatter plots with associated r values. Notice that, as r approaches 1, the array of dots converges towards a more linear pattern.

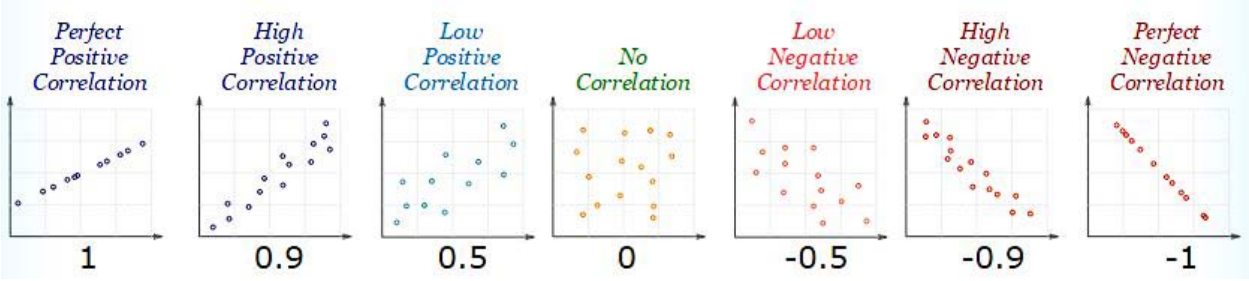


Image Credit: Correlation. (2020, December 10). AplusTopper website. Retrieved from https://www.aplustopper.com/correlation/

Here's how to interpret the correlation coefficient r:

- A positive value of r tells us that, as x increases, the y value will increase (and vice-versa). A positive r value indicates a positive correlation.
- A negative value of r tells us that, as x increases, y will decrease (and vice-versa). A negative r value indicates a negative correlation.
- It is also worth noting that the sign of r is always the same as the sign of the slope b of the best fit line.

Please keep in mind that correlation doesn't mean causation. While correlation can tell us that there is *some kind* of relationship between the variables, it doesn't identify that relationship and it doesn't typically mean that the relationship holds true for *all* circumstances. For example, humans do typically gain weight with age, but that doesn't mean that age causes weight gain.

Testing the Significance of the Correlation Coefficient

While it is true that the correlation coefficient r tells us something about the strength and direction of the linear relationship between x and y, the reliability of this linear model also depends upon the size of the set of data points in the sample. Therefore, to fully gain an understanding of the relationship illustrated by r, we also need to consider the sample size n.

Since we don't know the correlation coefficient for the entire population ρ, we use r as the sample correlation coefficient. The null hypothesis will be that there is no relationship between x and y represented as $H_o: \rho = 0$. The alternative hypothesis will there is a relationship between x and y represented as $H_a: \rho \neq 0$.

At determined level of significance (usually $\alpha = .05$),a test statistic (based on the t-distribution) is calculated and a p-value is determined. Just like the other hypothesis tests you've learned so far, we use the p-value to make the decision to *reject* or *fail to reject* the null hypothesis.

The Coefficient of Determination r^2

The coefficient of determination r^2 is simply the square of the correlation coefficient r. The coefficient of determination is typically reported as a percentage that represents the amount of variability in the dependent variable y that can be explained by variation in the independent variable x using the least squares regression line. Conversely, $1-r^2$ expressed as a percentage shows the variability in y that is not explained by x using the regression line. $1-r^2$ can be visualized as the scattering in the data points about the regression line. The more the pattern is scattered, the higher the value of $1-r^2$.

Example Problems

Example 1: The following data set displays contributions made during the first ten years of Cityville's annual zoo luncheon *"Eat with the Animals."* Perform a regression analysis to determine the linear equation that represents the relationship between year and contributions. Calculate the correlation coefficient and the coefficient of determination. Note that the event was not held in years 2, 8, and 9 due to budget restrictions. Project the contribution amount that could be expected in year 11.

Year	Contributions (dollars)
1	500
3	300
5	100
6	80
7	75
10	50

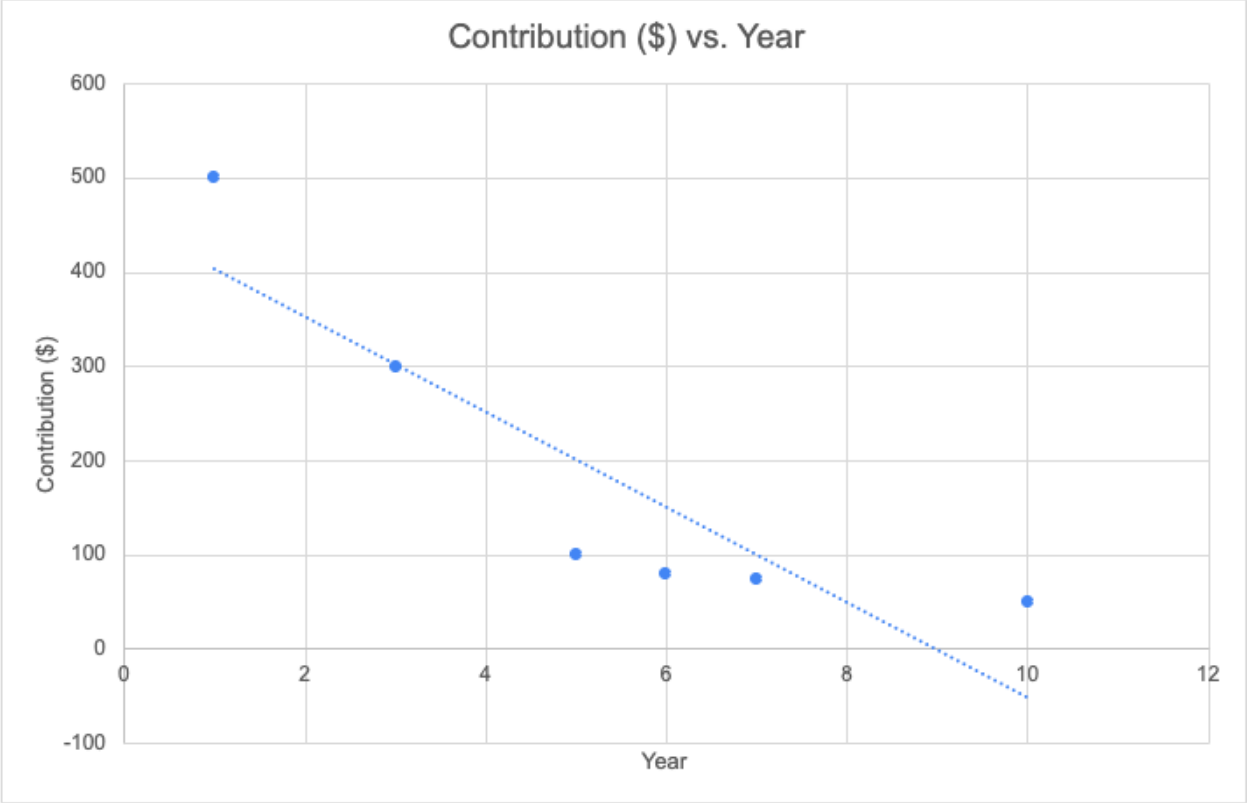
Use the spreadsheet provided in Math 201 Week 5, we calculate the following linear regression values:

Linear Regression Calculations	
Slope	-50.44
y-intercept	453.18
correlation coefficient r	-0.8829
coefficient of determination r^2	0.7795

Our correlation coefficient r = -0.8829 indicates that there is a strong negative correlation between year and contribution level. The coefficient of determination r^2 = 0.7795 indicates that 77.95% of the variability in contributions can be explained by the variable year. This verifies the assumption that this fundraising event has become less effective in collecting contributions over time. In fact, with a slope of -50.44, it means that each year the event was held, it lost $50.44 per year. The linear equation (or regression equation) for the best fit line is:

Contributions (in dollars) = 453.18 - 50.44*Year

Here is the scatter plot for this scenario showing the linear regression best fit-line.



If the event were held in year 11, the zoo could expect to receive the following in contributions:

Contributions (in dollars) = 453.18 - 50.44*(11) = -101.66

The prediction is that, during year 11, the event will lose $101.66. Time to reconsider the efficacy of this fund raising event. Of course, this is a very simple example. A more detailed analysis of the situation would also need to consider other factors such as expenses, time of year, weather, and competition with other venues.

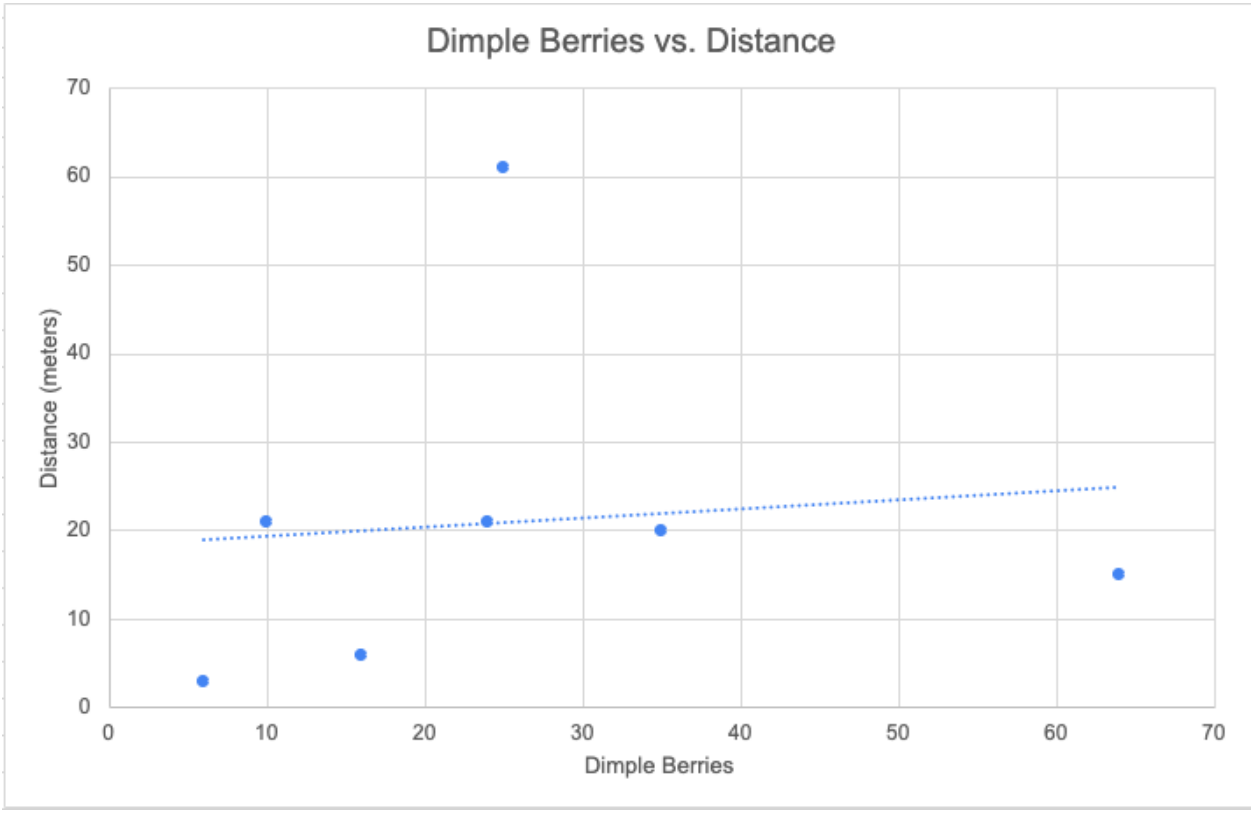
Example 2: Researchers observing the activity of *griffagongs* wanted to see if there is a correlation between the number of dimple berries they consumed and the distance they would travel to pick them (in meters). The following table represents the data they collected. Draw a scatter plot and regression line. Determine the correlation coefficient r and the coefficient of determination r^2.

# Dimple Berries	Distance to Dimple Berry Bush from Home Tree (meters)
24	21
6	3
16	6
64	15
10	21
25	61
35	20

Linear Regression Calculations	
Slope	0.10
y-intercept	18.38
correlation coefficient r	0.1042
coefficient of determination r^2	0.0109

According to our calculations, the correlation coefficient r = 0.1042 indicates a very weak positive correlation between the variables. This indicates that there isn't much of a relationship between the number of dimple berries consumed and the distance that *griffagongs* travel to collect them. This interpretation is further supported by the coefficient of determination r^2 = 0.0109 which indicates that only 1.09% of the variability in distance traveled is explained by the number of berries consumed.

Here's the scatter plot with the linear regression line.



This scatter plot clearly shows that there is no discernable pattern in the data further supporting the results reported earlier. Notice that the linear regression best-fit line is nearly horizontal. This is a further clue that there exists no correlation between the variables.

Example 3: Researchers are interested in determining the correlation between the height of Dmeubo bushes (in centimeters) and number of berries harvested from the bush. The following table of data was collected to answer this question. Create a scatter plot and draw a regression line. Determine the correlation coefficient r and the coefficient of determination r^2.

Bush Height (cm)	# Berries Harvested
68	7
74	4
82	8
88	10
93	11
99	9
101	13

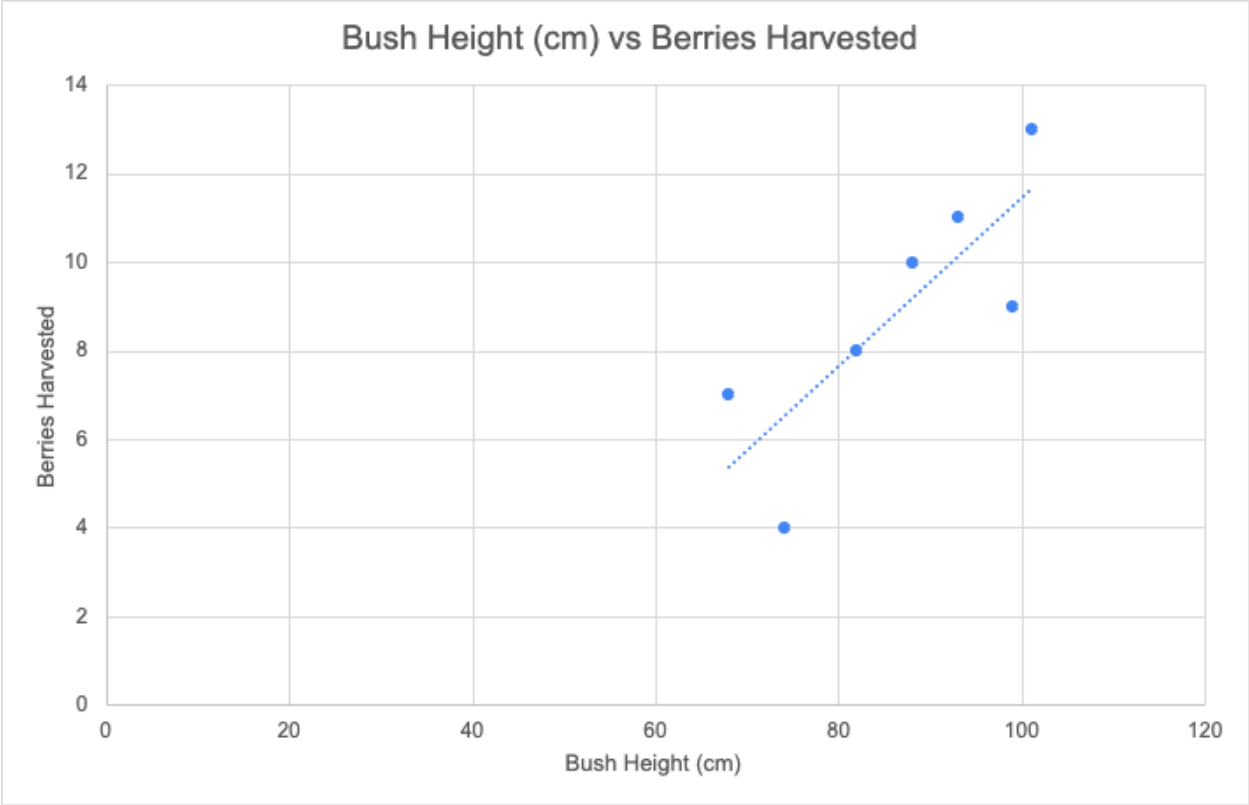
Linear Regression Calculations	
Slope	0.19
y-intercept	-7.54
correlation coefficient r	0.8114
coefficient of determination r^2	0.6583

Since the correlation coefficient r = 0.8114, there is a strong positive correlation between the bush height and number of berries harvested. The coefficient of determination r^2 = 0.6583 indicates that 65.83% of the variability in berries harvested is explained by bush height. Keep in mind that this is a very small set of data. A larger sample would help to reveal the true strength in the relationship between the variables.

Here is the linear regression equation for this relationship:

Berries Harvested = -7.54 + 0.19*Bush Height

The slope, 0.19, indicates that for every centimeter of bush height, there is a 0.19 increase in the number of available berries to harvest. The scatter plot with linear regression best-fit line follows:


Bush Height (cm) vs Berries Harvested
Berries Harvested
Bush Height (cm)
0
2
4
6
8
10
12
14
0
20
40
60
80
100
120