**Note to the Teacher:**

Attached here are additional teacher references for this lesson that can be used to give more information to the students such as definitions, deeper understanding, or hands on activities for students to complete. For example, if you would like students to generate correlation coefficient and linear regression model by hand. Or, these references can also be used as way to edit the lesson to your liking. This information can also be used as a teacher reference when teaching students the concepts of correlation and regression.

**References Listed:**

Page 1: Explanation of the Coefficient Correlation of Determination, R²

#### Page 2: Explanation of Correlation Coefficient, r

Page 3: How to Calculate Correlation Coefficient by Hand

**Page 4: Further Explaining Correlation**

**Page 5: The Regression Line**

**Page 6: How to Find the Regression Line By Hand**

**Page 7: How to Find the Regression Line By Hand Cont.**

**Page 8: Video Instructions, which can be Used to Teach Students How to:**

* Enter paired data into lists
* Produce a scatter plot
* Creating a regression equation for a set of data

**Explanation of the Coefficient Correlation of Determination, R²:**

#### Definition:

The last number to look at (third on the screen) is R², the **coefficient of determination**. (The calculator displays r², but the capital letter is standard notation.) R² measures the quality of the regression line as a means of predicting ŷ from x: **the closer R² is to 1, the better the line.** Another way to look at it is that R² measures **how much of the total variation in y is predicted by the line**.

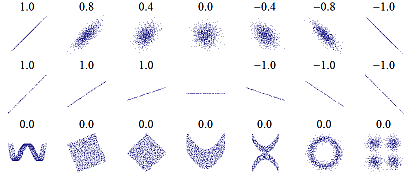
In this case R² is about 0.88, so your interpretation is “about 88% of the variation in distance traveled is associated with variation in club-head speed.” Statisticians say that R² tells you how much of the variation in y is “explained” by variation in x, but if you use that word remember that it means a numerical association, not necessarily a cause-and-effect explanation. It’s best to stick with “associated” unless you have done an experiment to show that there is cause and effect.

There’s a subtle difference between r and R², so keep your interpretations straight. r talks about the strength of the association between the variables; R² talks about what part of the variation in the y variable is associated with variation in the x variable, and how well the line predicts y from x. Don’t use any form of the word “correlated” when interpreting R².

Only linear regression will have a correlation coefficient r, but any type of regression — fitting any line or curve to a set of data points — will have a coefficient of determination R² that tells you how well the regression equation predicts y from the independent variable(s). Steve Simon (1999b) gives an example for non-linear regression in [R-squared](http://www.tc3.edu/instruct/sbrown/swt/sources.htm#so_Simon1999b) [see “Sources Used” at end of book].

In straight-line regression, R² is the square of r, so if you want a formula just [compute r](http://www.tc3.edu/instruct/sbrown/swt/chap04.htm#c04_r_formula) and square the result.

#### Explanation of Correlation Coefficient, r



“Several sets of (x,y) [pairs], with the correlation coefficient for each set. Note that correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom).”   
source: [Correlation and Dependence](http://www.tc3.edu/instruct/sbrown/swt/sources.htm#so_Correlation2014) [see “Sources Used” at end of book]

Look first at r, the **coefficient of linear correlation**. **r can range from −1 to +1** and measures the strength of the association between x and y. A positive correlation or **positive association** means that y tends to increase as x increases, and a negative correlation or **negative association** means that y tends to decrease as x increases. The closer r is to 1 or −1, the stronger the association. We usually round r to two decimal places.

[Karl Pearson](http://www.tc3.edu/instruct/sbrown/swt/bignames.htm#bign_PearsonK) developed the formula for the linear correlation coefficient in 1896. The symbol r is due to Sir Francis Galton in 1888.

For real-world data, 0.94 is a pretty strong correlation. But you might wonder whether there’s actually a general association between club-head speed and distance traveled, as opposed to just the correlation that you see in this sample. [Decision Points for Correlation Coefficient](http://www.tc3.edu/instruct/sbrown/swt/chap04.htm#c04_decpts_root), later in this chapter, shows you how to answer that question.

*Why is r positive when data points trend up to the right and negative when they trend down to the right? The product (x−x̅)(y−y̅) explains this. When points trend up to the right, most are in the lower left and upper right quadrants of the plot. In the lower left, x and y are both below average, x−x̅ and y−y̅ are both negative, and the product is positive. In the upper right, x and y are both above average, x−x̅ and y−y̅ are both positive, and again the product is positive. The product is positive for most points, and therefore r is positive when the trend is up to the right.*

*On the other hand, if the data trend down to the right, most points are in the upper left (where x is below average and y is above average, x−x̅ is negative, y−y̅ is positive, and the product is negative) and the lower right (where x−x̅ is positive, y−y̅ is negative, and the product is also negative.) Since the product is negative for most points, r is negative when data trend down to the right.*

Reference: <http://www.tc3.edu/instruct/sbrown/swt/chap04.htm#c04_regress_root>

**How to Calculate Correlation Coefficient by Hand**

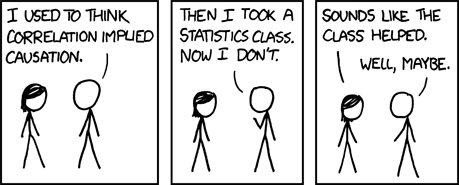
We will begin by listing the steps to the calculation of the correlation coefficient. The data we are working with are [paired data](http://statistics.about.com/od/Glossary/a/What-Is-Paired-Data.htm), each pair of which will be denoted by (*xi,yi*). Found on the data sheet above. Then, use a different (blank) sheet of paper and your graphing calculator to find the correlation coefficient.

1. We begin with a few preliminary calculations. The quantities from these calculations will be used in subsequent steps of our calculation of *r*:
   * 1. The correlation coefficient r, denotes its [value](http://www.businessdictionary.com/definition/value.html) varies between -1 and 1 : 1 [means](http://www.businessdictionary.com/definition/mean.html) [perfect](http://www.businessdictionary.com/definition/perfect.html) correlation, 0 means no correlation, positive [values](http://www.businessdictionary.com/definition/values.html) means the [relationship](http://www.businessdictionary.com/definition/relationship.html) is positive (when one goes up so does the other), negative values mean the relationship is negative (when one goes up the other goes down).
2. Calculate x̄, the [mean](http://statistics.about.com/od/HelpandTutorials/a/Ways-To-Find-The-Average.htm) of all of the first coordinates of the data *xi*.
3. Calculate ȳ, the mean of all of the second coordinates of the data *yi*.
   * 1. To calculate the mean: add up all the numbers, then divide by how many numbers there are.
4. Calculate *s x* the sample [standard deviation](http://statistics.about.com/od/HelpandTutorials/a/How-To-Calculate-A-Standard-Deviation.htm) of all of the first coordinates of the data *xi*. It measures how spread out of the numbers fall away from the standard deviation.
5. Calculate *s y* the sample standard deviation of all of the second coordinates of the data *yi*.
   * 1. To calculate the standard deviation: take the square root of the Variance.
6. Use the formula *(zx)i* = (*xi* – x̄) / *s x* and calculate a standardized value for each *xi*.
7. Use the formula *(zy)i* = (*yi* – ȳ) / *s y* and calculate a standardized value for each *yi*.
8. Multiply corresponding standardized values: *(zx)i(zy)i*
9. Add the products from the last step together.
10. Divide the sum from the previous step by n – 1, where n is the total number of points in our set of paired data.
11. The result of all of this is the correlation coefficient *r*.

**Further Explaining Correlation:**

**Be careful in your interpretation!** No matter how strong your r might be, say that changes in the y variable are associated with changes in the x variable, not “caused by” it. **Correlation is not causation** is your mantra.

It’s easy to think of associations where there is no cause. For example, if you make a scatterplot of US cities with x as number of books in the public library and y as number of murders, you’ll see a positive association: number of murders tends to be higher in cities with more library books. Does that mean that reading causes people to commit murder, or that murderers read more than other people? Of course not! There is a **lurking variable** here: population of the city.

When you have a positive or negative association, there are **four possibilities**: x might cause changes in y, y might cause changes in x, lurking variables might cause changes in both, or it could just be coincidence, a random sample that happens to show a strong association even though the population does not.

used by permission; source: <http://xkcd.com/552/> (accessed 2014-09-15)

If correlation is not causation, then how can we establish causation? For example, how do we know that smoking causes lung cancer in humans? Obviously we can’t perform an experiment, for ethical reasons. [Sir Austin Bradford Hill](http://www.tc3.edu/instruct/sbrown/swt/bignames.htm#bign_BradfordHill) laid down nine criteria for establishing causation in a 1965 paper, [The Environment and Disease: Association or Causation?](http://www.tc3.edu/instruct/sbrown/swt/sources.htm#so_BradfordHill1965) [see “Sources Used” at end of book] Short summaries of the “Bradford Hill criteria” are many places on the Web, including Steve Simon’s (2000a) [Causation](http://www.tc3.edu/instruct/sbrown/swt/sources.htm#so_Simon2000a) [see “Sources Used” at end of book].

Reference: <http://www.tc3.edu/instruct/sbrown/swt/chap04.htm#c04_regress_root>

**The Regression Line:**

**Definition:**

Write the equation of the line using ŷ (“y-hat”), not y, to indicate that this is a prediction. b is the **y intercept**, and a is the **slope**. Round both of them to four decimal places, and write the equation of the line as

ŷ = 3.1661x − 55.7966

(Don’t write 3.1661x + −55.7966.)

These numbers can be interpreted pretty easily. Business majors will recognize them as intercept = fixed cost and slope = variable cost, but you can interpret them in non-business contexts just as well.

The **slope**, a or b1 or m, tells **how much ŷ increases or decreases for a one-unit increase in x**. In this case, your interpretation is “the ball travels about an extra 3.17 yards when the club speed is 1 mph greater.” The slope and the correlation coefficient always have the same sign. (A negative slope would mean that y decreases that many units for every one unit increase in x.)

The **intercept**, b or b0, says where the regression line crosses the y axis: it’s the **value of ŷ when x is 0**. Be careful! **The y intercept may or may not be meaningful.** In this case, a club-head speed of zero is not meaningful. In general, when the measured x values don’t include 0 or don’t at least come pretty close to it, you can’t assign a real-world interpretation to the intercept. In this case you’d say something like “the intercept of −55.7966 has no physical interpretation because you can’t hit a golf ball at 0 mph.

Here’s an example where the y intercept does have a physical meaning. Suppose you measure the gross weight of a UPS truck (y) with various numbers of packages (x) in it, and you get the regression equation ŷ = 2.17x+2463. The slope, 2.17, is the average weight per package, and the y intercept, 2463, is the weight of the empty truck.

*The slope (a or m or b1) and y intercept (b or b0) of the regression line can be calculated from formulas, if you have a lot of time on your hands:*



*For the meaning of* ∑*, see* [*∑ Means Add ’em Up*](http://www.tc3.edu/instruct/sbrown/swt/chap01.htm#c01_BigSigma) *in Chapter 1.*

*Traditionally, calculus is used to come up with those equations, but all that’s really necessary is some algebra. See* [*Least Squares — the Gory Details*](http://www.tc3.edu/instruct/sbrown/stat/leastsq.htm) *if you’d like to know more.*

*The second formula for the slope is kind of neat because it connects the slope, the correlation coefficient, and the SD of the two variables.*

Reference: <http://www.tc3.edu/instruct/sbrown/swt/chap04.htm#c04_regress_root>

**How to Find the Regression Line By Hand**

**Directions: Now that you have found the coefficient correlation, take the sum of the x’s added up, the sum of the y’s, the sum of the (x’s times y’s) and the sum of the x and y squares. The table below will guide you. Please use your graphing calculator to sum your values.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subject** | **TV Watching (x, hours)** | **Physical Activity (y, hours)** | **x \* y** | **x^2** | **y^2** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| **Sum (Totals)** |  |  |  |  |  |

**How to Find the Regression Line By Hand Cont.**

**Step 1:**

**Now, feel free to write your sum (totals below). Reminder “**Σ” stands for sigma which is the eighteenth letter of the Greek alphabet, carrying the “s” sound meaning to sum.

Σx =

Σy =

Σxy =

Σx^2 =

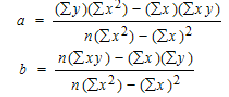
Σy^2 =

\*n is the sample size

**Step 2:**

Use the following equations to find a and b. In the linear regression equation:

y = ax +b.



a (slope) =

b (y-intercept) =

**Step 3:**

Now that you have the values for slope and y intercept, write the equation for the line of best fit, which best represents your data below:

y = \_\_\_\_\_\_\_\_x + \_\_\_\_\_\_\_\_

**Video Instructions, which can be Used to Teach Students How to:**

* Enter paired data into lists
* Produce a scatter plot
* Creating a regression equation for a set of data
  + Located at: <http://www.atomiclearning.com/ti_84>
    - Under C. Working with Lists