

CHALLENGING MISCONCEPTIONS ABOUT THE COEFFICIENT OF DETERMINATION R^2 WITH A DYNAMIC GEOGEBRA APPLET

Michael S. Meagher¹, Michael Todd Edwards² and S. Asli Özgün-Koca³

¹Brooklyn College–Brooklyn, New York

²Miami University–Oxford, Ohio

³Wayne State University–Detroit, Michigan

Abstract

The authors present an applet, R^2 Explorer, that can be used to explore how the value of R^2 is related to data points and to challenge the misconception that R^2 measures “percentage of fit” of a regression line.

Keywords: linear regression, coefficient of determination, multiple linked representations

1 INTRODUCTION

More than ever, technology is important in teaching and learning school mathematics. The past 20 years have seen breathtaking technological developments—from the widespread adoption of global positioning systems and smartphones to advances in artificial intelligence—that have implications for how we teach and learn mathematics.

More than a decade ago, we wrote an article highlighting the use of technology—specifically, the *TI-Nspire* graphing calculator—to explore R^2 , the coefficient of determination (Edwards et al., 2008). Quoting *Principles and Standards for School Mathematics* (NCTM, 2000), we noted the importance of multiple linked representations in students’ conceptual development:

Different representations often illuminate different aspects of a complex concept or relationship . . . the importance of using multiple representations should be emphasized throughout students’ mathematical education (p. 68).

Our article walked readers through the steps to construct a dynamic statistics sketch with *TI-Nspire*—one that linked a graph of points with a line of best fit and the calculation of R^2 . Once readers completed the construction, they used the sketch to explore a popular misconception—namely, that R^2 is a measure of the quality of the line of best fit (rather than a measure of how much variation in y can be accounted for by variation in x). The article led readers to arrange points that were close to horizontal and, therefore, had a “good fit” but a low R^2 value since the variation in y was too small to account for variation in x .

Although the article calls on classroom teachers to focus on students' conceptual understanding through multiple linked representations, and in many ways, it does, literally four of the article's five pages are devoted to construction steps with *TI-Nspire*. This was necessary and indeed encouraged by the journal, owing to the new technology. Unfortunately, as we revisited the article, it was apparent that the many construction steps distracted readers from the conceptual aspects of the task.

This new article is, in part, a response to our earlier work. In the following pages, we revisit R^2 with an interactive GeoGebra applet, *R² Explorer*. Technological advances since the publication of our earlier article have made it possible to provide teachers with a pre-made applet, freeing them to focus on a conceptual approach to the study of R^2 without button-pressing distractions. We've also included new features in *R² Explorer*—including dynamic, color-coded text and intermediate calculations for variance, covariance, and R^2 itself. These features encourage students to make connections between data points and the values of variance, covariance, and R^2 as they gain a deeper understanding of what effect arrangements of points have on the formulae and their values.

In the following pages, we highlight the interactive and visual nature of *R² Explorer*. Through a series of short examples, we illustrate how students can manipulate data points and observe real-time changes in various statistical values. The remainder of this paper is broken into 3 sections. In *Section 2*, we briefly overview the applet and its components. *Section 3* provides an in-depth, hands-on exploration of the applet. *Section 4* provides a summary of key findings and insights, emphasizing ways the applet may enhance student understanding of statistics by analyzing intermediate calculations.

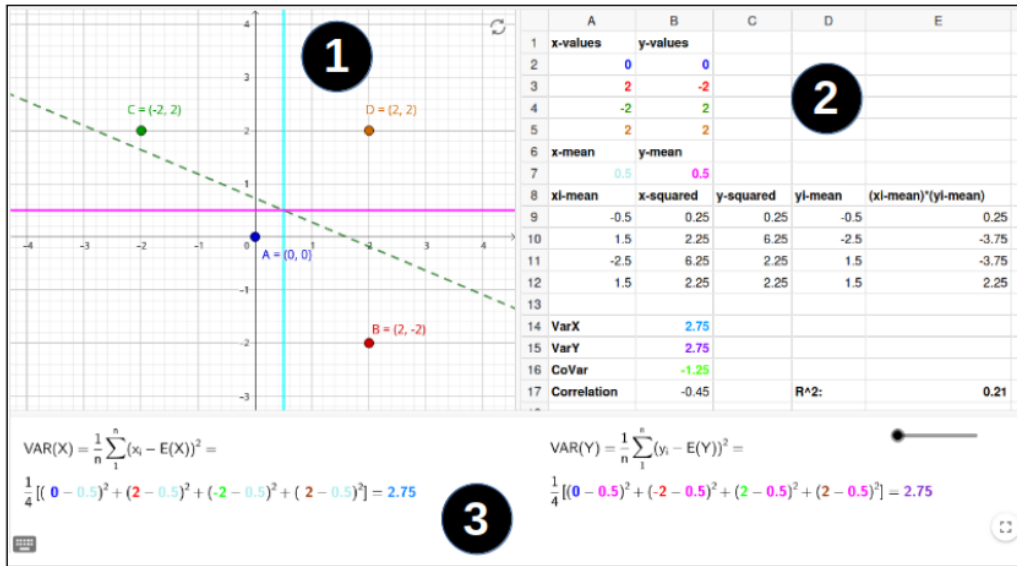
2 INTRODUCING R^2 EXPLORER

Unfortunately, many of the student challenges we reported in our original article (Edwards et al., 2008) persist into the present. Students continue to misinterpret the value of R^2 as a “percentage of fit” (Edwards et al., 2008) rather than the “proportion of the variance” in the response variable that can be explained by the predictor variable in the regression model (Statology, 2023). Since the values of R^2 range between 0 and 1, this misconception is difficult to eradicate.

We developed *R² Explorer* to support our students in developing deeper, more nuanced understandings of linear regression and R^2 . This section provides a comprehensive overview of the applet's functionality, emphasizing its interactive features and the engagement it fosters among students. Figure 1 provides a screenshot of the applet.

As Figure 1 suggests, the applet consists of a split-pane interface, with the top pane featuring a Graph View and a Spreadsheet View side-by-side (i.e., Regions 1 and 2, respectively). A bottom Equation Pane (i.e., Region 3) provides students with a series of general formulae (i.e., variance, covariance, and coefficient of determination, R^2) accessed with a slider, with an instantiation of each of the three formulas based on current values displayed in Region 2.

Figure 1. Layout of the R^2 Explorer applet.

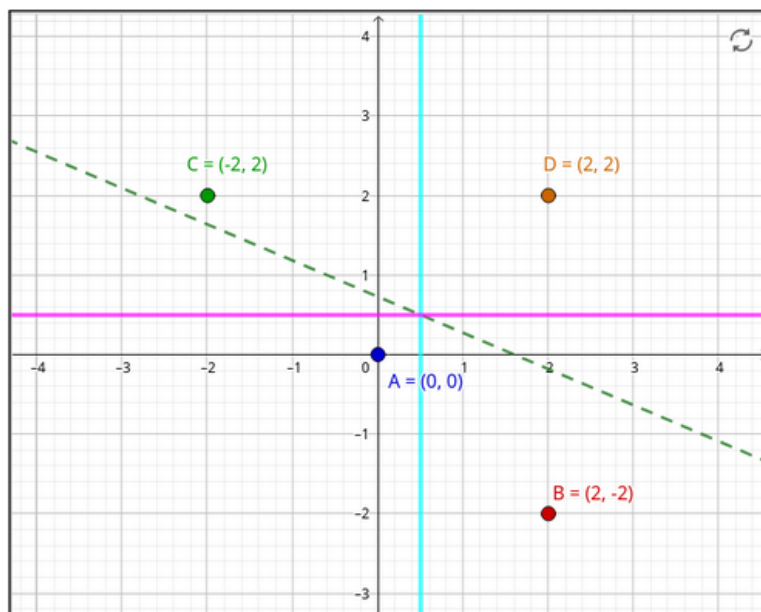


For now, don't worry too much about the colored text. We'll discuss its significance in more detail later in the paper. We begin with a discussion of the basic features of each pane in the remainder of this section and provide a hands-on tour of the applet in *Section 3, A Hands-on Tour of the Applet*.

Graph View

As Figure 2 illustrates, the graph in R^2 Explorer includes: (a) 4 data points— A , B , C , and D ; (b) a coordinate plane with x - and y -axes; (c) a least squares regression line dotted **green**; (d) a vertical **cyan** line representing the mean of the x -coordinates of the data points; and (e) a horizontal **magenta** line representing the mean of the y -coordinates of the data points.

Figure 2. Graph View.



Spreadsheet View

Consider the Spreadsheet View shown in Figure 3.

Figure 3. Spreadsheet View.

| | A | B | C | D | E | |
|----|-------------|-----------|-----------|---------|---------------------|-------|
| 1 | x-values | y-values | | | | |
| 2 | | 0 | | | | |
| 3 | | 2 | | | | |
| 4 | | -2 | | | | |
| 5 | | 2 | | | | |
| 6 | x-mean | y-mean | | | | |
| 7 | | 0.5 | | | | |
| 8 | xi-mean | x-squared | y-squared | yi-mean | (xi-mean)*(yi-mean) | |
| 9 | | -0.5 | 0.25 | 0.25 | -0.5 | 0.25 |
| 10 | | 1.5 | 2.25 | 6.25 | -2.5 | -3.75 |
| 11 | | -2.5 | 6.25 | 2.25 | 1.5 | -3.75 |
| 12 | | 1.5 | 2.25 | 2.25 | 1.5 | 2.25 |
| 13 | | | | | | |
| 14 | VarX | 2.75 | | | | |
| 15 | VarY | 2.75 | | | | |
| 16 | CoVar | -1.25 | | | | |
| 17 | Correlation | -0.45 | | R^2: | 0.21 | |
| 18 | | | | | | |

Next, note the following values contained within the sheet:

- The x and y coordinates of each data point, along with information such as the mean of all x -coordinates (labeled as “ x -mean”) and the mean of all y -coordinates (labeled as “ y -mean”);
- Intermediate calculations, including (a) differences between each x -coordinate and the mean of all x -coordinates (labeled as “ xi -mean”); (b) the differences between each y -coordinate and the mean of all y -coordinates (labeled as “ yi -mean”); (c) the squares of these, labeled as x -squared and y -squared, respectively; and (d) the product of these differences (labeled as “(xi-mean)*(yi-mean)”);
- The variance of x ($\text{Var}(X)$), variance of y ($\text{Var}(Y)$), covariance ($\text{Cov}(X, Y)$), and the coefficient of determination (R^2).

Equation Pane

Next, consider the Equation Pane illustrated in Figure 4.

Figure 4. The Equation Pane.

| | |
|---|---|
| $\text{VAR}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))^2 =$ $\frac{1}{4} [(0 - 0.5)^2 + (2 - 0.5)^2 + (-2 - 0.5)^2 + (2 - 0.5)^2] = 2.75$ | $\text{VAR}(Y) = \frac{1}{n} \sum_{i=1}^n (y_i - E(Y))^2 =$ $\frac{1}{4} [(0 - 0.5)^2 + (-2 - 0.5)^2 + (2 - 0.5)^2 + (2 - 0.5)^2] = 2.75$ |
|---|---|

The slider on the right controls the content displayed on the left. Dragging the slider, students choose from 3 possible formulas:

- Variance: $\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n [x_i - E(X)]^2$ and $\text{Var}(Y) = \frac{1}{n} \sum_{i=1}^n [y_i - E(Y)]^2$
- Covariance: $\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$
- Coefficient of Determination: $R^2 = \frac{\text{Cov}(X,Y) \cdot \text{Cov}(X,Y)}{\text{Var}(X) \cdot \text{Var}(Y)}$

Dynamic Features of the Applet

As a GeoGebra applet, *R² Explorer* is dynamic. In other words, as students drag data points to new locations within the grid, values in the Spreadsheet View and Equation Pane are updated automatically, reflecting the new position of the points and the new coordinate values. As students interact with the applet, we encourage them to look for relationships among values in the Spreadsheet View, Equation Pane, and the plotted data points in the Graph View—an idea we explore in more detail in the next section, *A Hands-On Tour of the Applet*.

3 A HANDS-ON TOUR OF THE APPLET

Rather than simply telling our readers about *R² Explorer*, we believe engaging you with the applet as you read is important. With this in mind, access a copy of the sketch at <https://www.geogebra.org/m/v2tukpc5>. Once you're there, look at the Graph View, Spreadsheet View, and Equation Pane. *What do you see? What do you notice? What do you wonder?* (NCTM, 2023). The following paragraphs share some observations in short, hands-on explorations.

Exploring Means

Let's begin by looking at the values of *x*-mean and *y*-mean within the spreadsheet (i.e., cells A7 and B7). Drag the points in Graphics View to the following locations: *A*(1, 0), *B*(2, 0), *C*(2, 0), and *D*(1, 0). Your graph should look similar to Figure 5. *What do you notice about the solid vertical line? What is the equation of this line? How are the locations of the plotted points related to the equation of the line?*

Figure 5. Graphics View with data points located at *A*(1, 0), *B*(2, 0), *C*(2, 0), and *D*(1, 0).

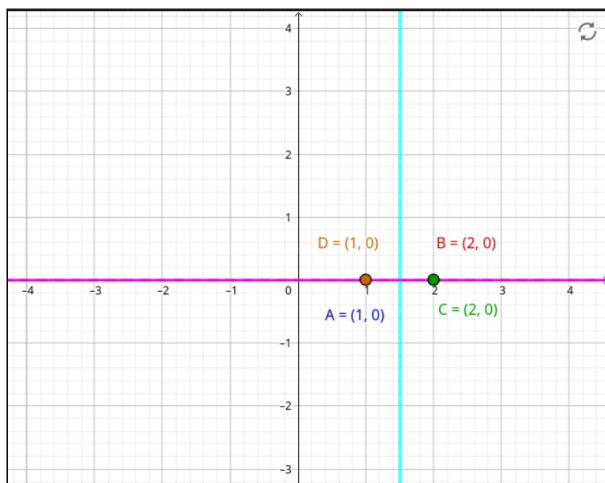
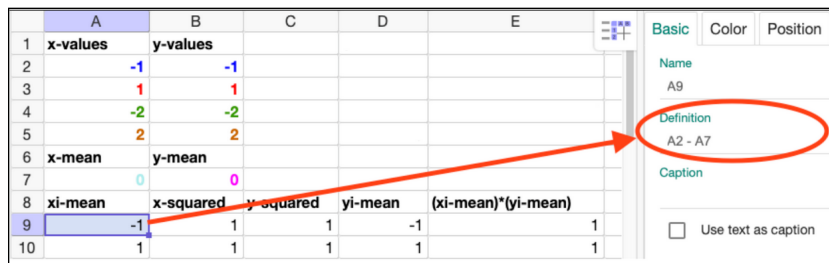


Figure 5 suggests that the solid vertical line intersects the x -axis at $x = 1.5$. How is that value related to the position of the two points at $x = 1$ and $x = 2$? After some conversation and experimentation with the applet, many students recognize that the vertical line represents the mean of the x -coordinates (i.e., $(1 + 1 + 2 + 2)/4 = 1.5$). Likewise, the horizontal line represents the mean of the y -coordinates. Drag data points around the grid until you are convinced this is true.

Exploring Expected Values

Next, note the values of x_i -mean and y_i -mean within the spreadsheet (i.e., cells A9–A12 and D9–D12, respectively). Right-click on cell A9 and select Object Properties from the contextual menu. What do you notice? (NCTM, 2023). The first x_i -mean is calculated by taking the difference of the first x -value (cell A2) and the mean of the four x -coordinates (cell A7). Figure 6 illustrates this calculation within the spreadsheet.

Figure 6. Calculation of x_i -mean for first data point, $(-1, -1)$.



Algebraically, the calculation for the i th x_i -mean is represented as

$$x_i - E(X),$$

where $E(X)$ represents the “expected value of x ” (i.e., the mean of x). The calculation for the i th y_i -mean is analogous, and is represented as

$$y_i - E(Y).$$

Exploring Variance

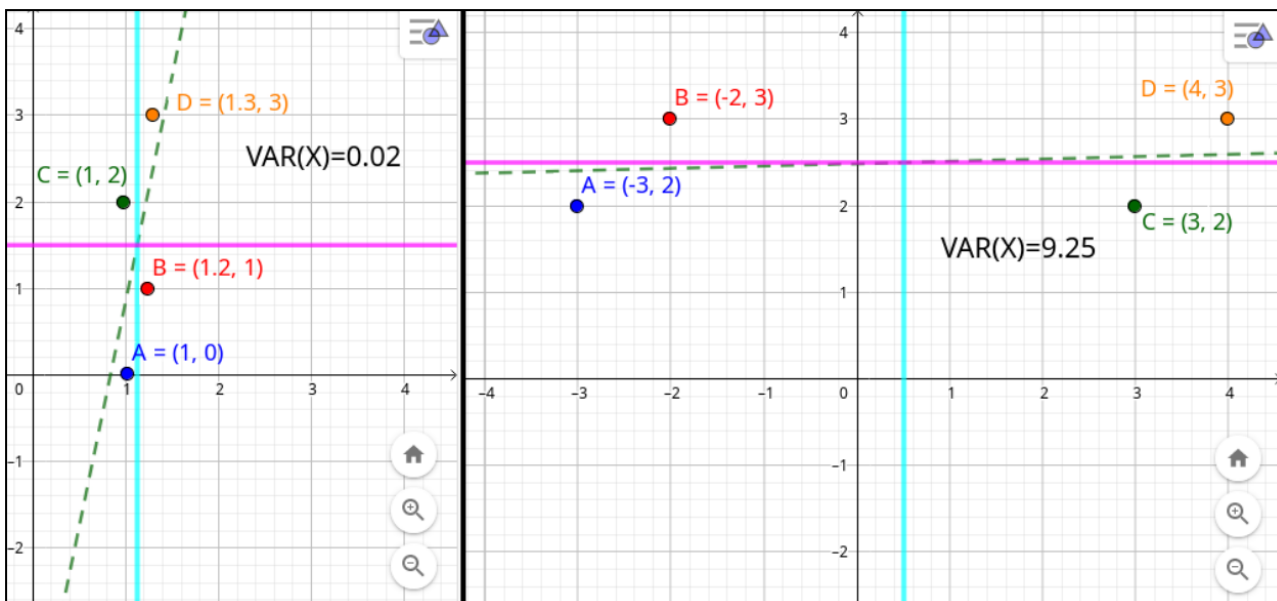
Averaging the squares of these differences across all data points (in this case, 4 points) gives us the formulas for the variance for X and Y , namely

$$\text{Var}(X) = \frac{1}{4} \sum_{i=1}^4 [x_i - E(X)]^2$$

$$\text{Var}(Y) = \frac{1}{4} \sum_{i=1}^4 [y_i - E(Y)]^2$$

We encourage students to move the points in the applet and explore the effect of the points’ position on the variance values. For example, what arrangements of points give small values for $\text{Var}(X)$ (say, less than 1) or large values of $\text{Var}(X)$ (say greater than 7)? Examples of such cases are illustrated in Figure 7.

Figure 7. (Left) Points arranged in a roughly vertical pattern exhibit small variation in x ($\text{Var}(X) = 0.02$); (Right) Points spread across a wide horizontal region exhibit large variation in x ($\text{Var}(X) = 9.25$).



Moreover, as you drag points, notice the values changing in the formula for $\text{Var}(X)$ in the applet's Equation Pane (i.e., Region 3). *What relationships are contributing, in arithmetical terms, to the size of the value of $\text{Var}(X)$?*

After some exploration, students should notice that when points are far away from the “vertical mean,” the values of $x - E(X)$ get large, and when those values are squared, they make for a large contribution to the value of $\text{Var}(X)$. Likewise, if the points get close to the mean for a particular coordinate, the values of $x - E(X)$ get small, and, in particular, if the difference is less than 1, then, upon squaring, the contribution to the variance is very small.

Here are some additional questions designed to encourage exploration of variance:

- What arrangements of points give small values for $\text{Var}(Y)$ or large values of $\text{Var}(Y)$?
- Can you find an arrangement of points that makes $\text{Var}(X)$ and/or $\text{Var}(Y)$ negative?
- Can you find an arrangement of points that gives a large value for $\text{Var}(X)$ and a small value for $\text{Var}(Y)$? What about a small value for $\text{Var}(X)$ and a large value for $\text{Var}(Y)$? What about small values for both? Large values for both?

Exploring these questions should help students see that “almost vertical” arrangements of points lead to small values for $\text{Var}(X)$ and “almost horizontal” arrangements of points lead to small values for $\text{Var}(Y)$. This foreshadows the effect that small variance values have on calculations of R^2 .

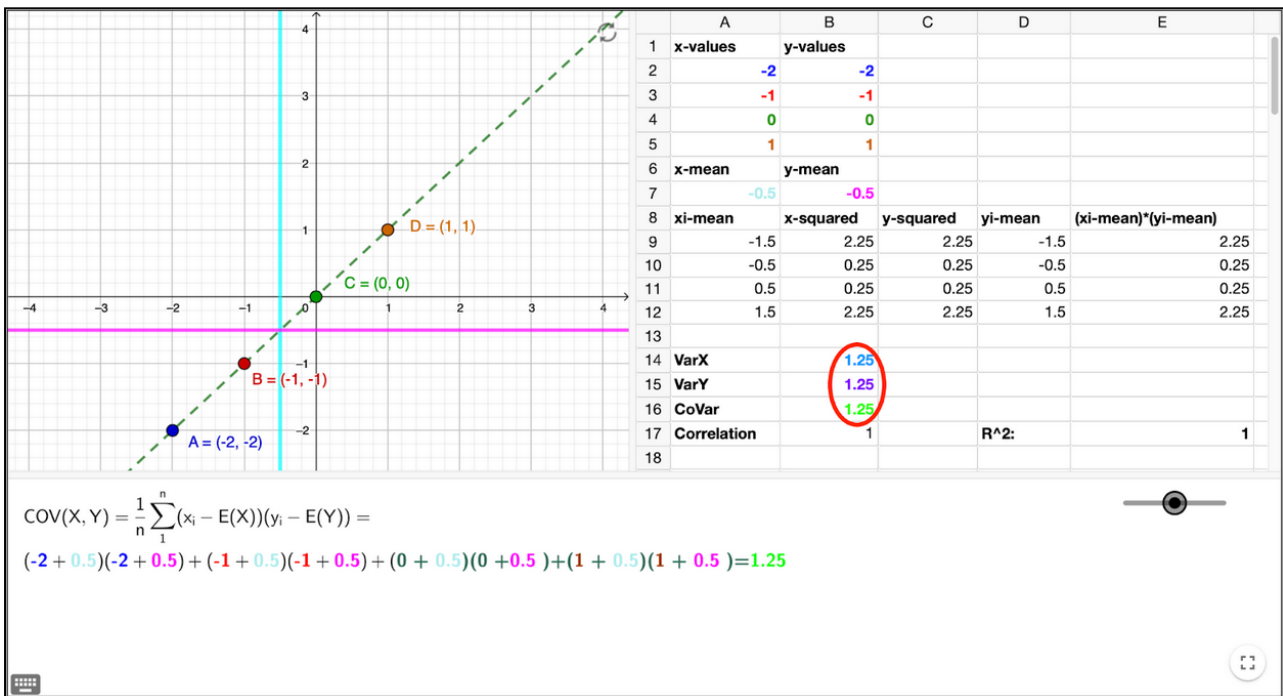
Exploring Covariance

Next, we explore covariance, which is found by averaging the product of $(x_i - E(X))(y_i - E(Y))$ for $1 \leq i \leq 4$. In other words,

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$$

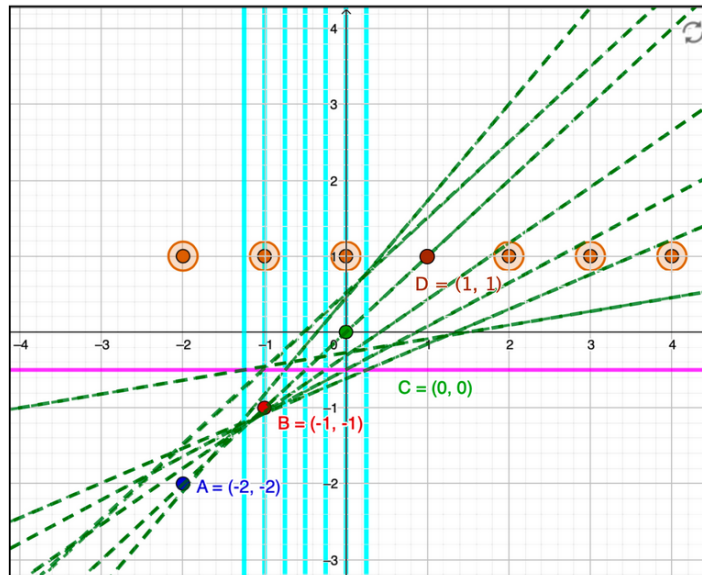
Covariance is a measure of how the plotted points vary together. To see this, move the slider to show the covariance formula (i.e., the left-most position) as shown in Figure 8. Drag the points in the graph to coordinates $A(-2, -2)$, $B(-1, -1)$, $C(0, 0)$, and $D(1, 1)$. Note that $\text{Var}(X) = \text{Var}(Y) = \text{Cov}(X, Y)$, as circled in Figure 8.

Figure 8. Arrangement of data points resulting in $\text{Var}(X) = \text{Var}(Y) = \text{Cov}(X, Y)$.



Next, move point D back and forth along the horizontal line $y = 1$ as suggested in Figure 9. *What is the effect on the values of $\text{Var}(X)$, $\text{Var}(Y)$, and $\text{Cov}(X, Y)$ as D is moved in this manner? How is this reflected arithmetically in the covariance formula?*

Figure 9. Point D at various positions along the horizontal line $y = 1$ with corresponding least squares fit lines.



Students should notice that the variance of Y is unchanged while the variance of X and the covariance change considerably. The y -value of the moving point remains constant while the x -value changes. How is this reflected in the formula for covariance? $E(X)$ is changing, so the values in **cyan** change in the formula. These changing values are underlined in Figure 10.

Figure 10. As students move the point $D(1, 1)$ back and forth along the horizontal line $y = 1$, the underlined values, $E(X)$, change.

$$\text{VAR}(X) = \frac{1}{n} \sum_1^n (x_i - E(X))^2 =$$

$$\frac{1}{4} [(0 - \underline{0.25})^2 + (4 - \underline{0.25})^2 + (-1 - \underline{0.25})^2 + (-2 - \underline{0.25})^2] = 5.19$$

Since the x -value of point D is changing, this change is also reflected in the covariance formula. Specifically, the underlined values in Figure 11 change.

Figure 11. As students move point D back and forth along the horizontal line $y = 1$, the underlined terms in $\text{Cov}(X, Y)$ change.

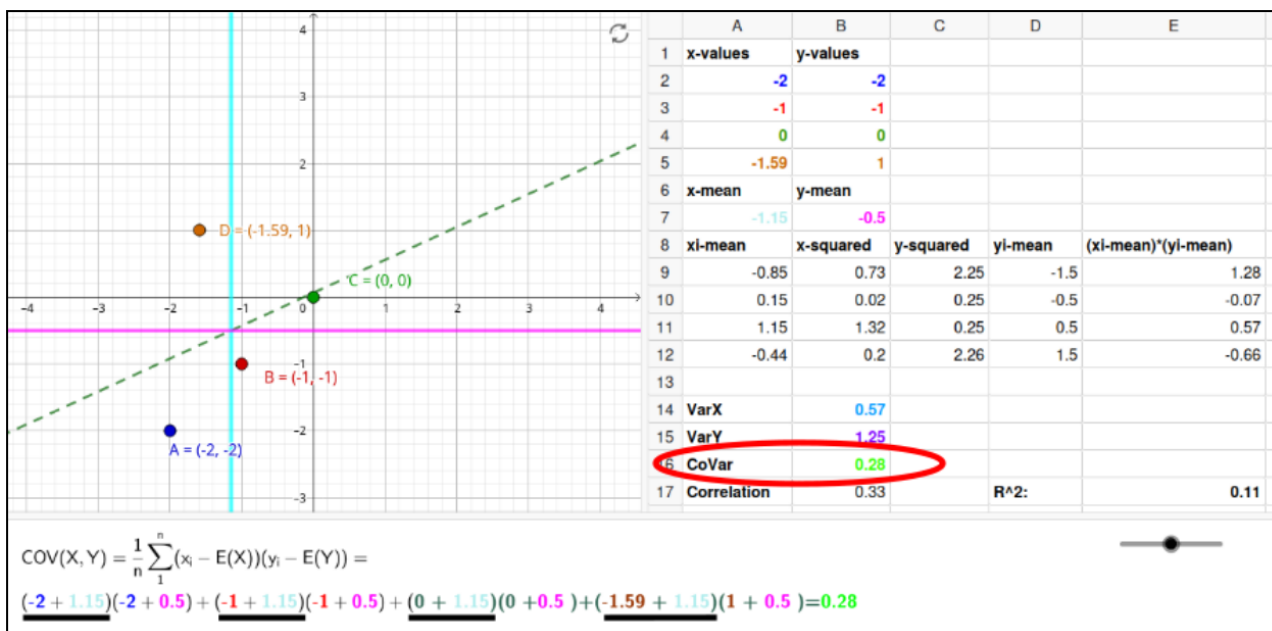
$$\text{COV}(X, Y) = \frac{1}{n} \sum_1^n (x_i - E(X))(y_i - E(Y)) =$$

$$\underline{(0 - 0.25)(0 + 0.5)} + \underline{(4 - 0.25)(1 + 0.5)} + \underline{(-1 - 0.25)(-1 + 0.5)} + \underline{(-2 - 0.25)(-2 + 0.5)} = 2.38$$

As point D is dragged to the right, the points get, on average, farther from the horizontal mean line (in **cyan**), and so the underlined values in Figure 11 get larger, and the covariance gets larger. Students observe a similar phenomenon if the point is dragged far to the left.

Now notice that if point D is dragged along the line $y = 1$ until its x -value falls between the x -values of points $A(-2, -2)$ and $C(0, 0)$, the points become, on average, close to the horizontal mean line (in cyan) and the values of the 1st, 3rd, 5th and 7th parenthetical calculations in $\text{Cov}(X, Y)$ get very small, resulting in a small covariance. Such an example is illustrated in Figure 12.

Figure 12. When point D is dragged along line $y = 1$ until it falls horizontally between points $A(-2, -1)$ and $C(0, 0)$, the resulting value of $\text{Cov}(X, Y)$ is relatively small since the underlined values in the Equation Pane are small.



If point D is returned to its original position $(1, 1)$ and then dragged up and down the vertical line $x = 1$, a similar argument can be made about the variance of Y and its effect on $\text{Cov}(X, Y)$. As students perform these steps, they see that more points clustered around a mean result in a smaller variance of one of the variables and smaller covariance.

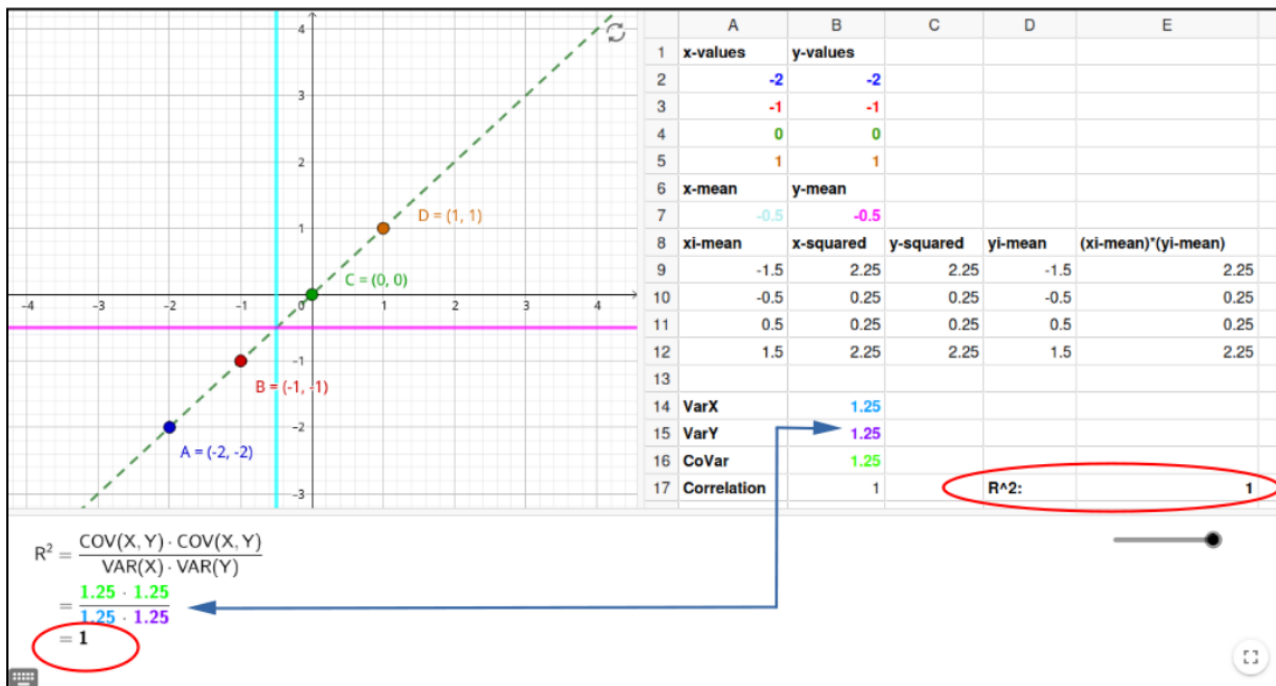
The final step of our exploration is to observe the effect of a small variance and the consequent small covariance on the calculation of R^2 .

3.1 Exploring R^2

With an understanding of variance and covariance from within the applet, students can begin exploring relationships between R^2 and positions of the plotted points within the graph.

Using the Equation Pane to Understand R^2 . Move your cursor to the Equation Pane at the bottom of R^2 Explorer. Next, move the data points to $A(-2, -2)$, $B(-1, 1)$, $C(0, 0)$, and $D(1, 1)$. Drag the slider to the right-most position to see the formula for R^2 in terms of variance and covariance as shown in Figure 13. The values in each formula correspond to values stored in the spreadsheet (i.e., the values in the Equation Pane and Spreadsheet are *dynamically linked*).

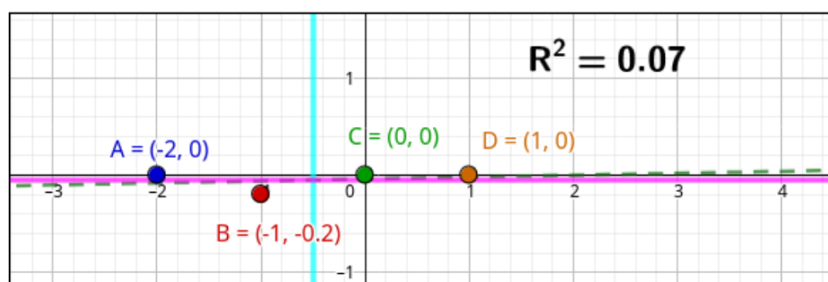
Figure 13. Plot of data points $A(-2, -2)$, $B(-1, -1)$, $C(0, 0)$, and $D(1, 1)$. The calculation of R^2 in the Equation Pane uses values stored in the Spreadsheet View.



In Figure 13, note that (a) $\text{Var}(X) = \text{Var}(Y) = \text{Cov}(X, Y)$, (b) the points are all on a straight line, and (c) the value of R^2 is 1. Too many students (and teachers) misinterpret this to mean that the line of best fit through the points is “100% perfect.” While that is true about the line of best fit, “percentage of fit” is not what R^2 is measuring.

Next, move the points to $A(-2, 0)$, $B(-1, -0.2)$, $C(0, 0)$, and $D(1, 0)$ as depicted in Figure 14.

Figure 14. Each of the points, $A(-2, 0)$, $B(-1, -0.2)$, $C(0, 0)$, and $D(1, 0)$, is close to the line of best fit. Based on this observation, students may assume that R^2 is close to 1.



Note that the points are all very close to the line of best fit, so students would think that the R^2 value should be close to 1. However, the value of R^2 is, in fact, 0.07. *How can this be explained?* Let’s look at the effect of the position of the points on the values in the formula for R^2 .

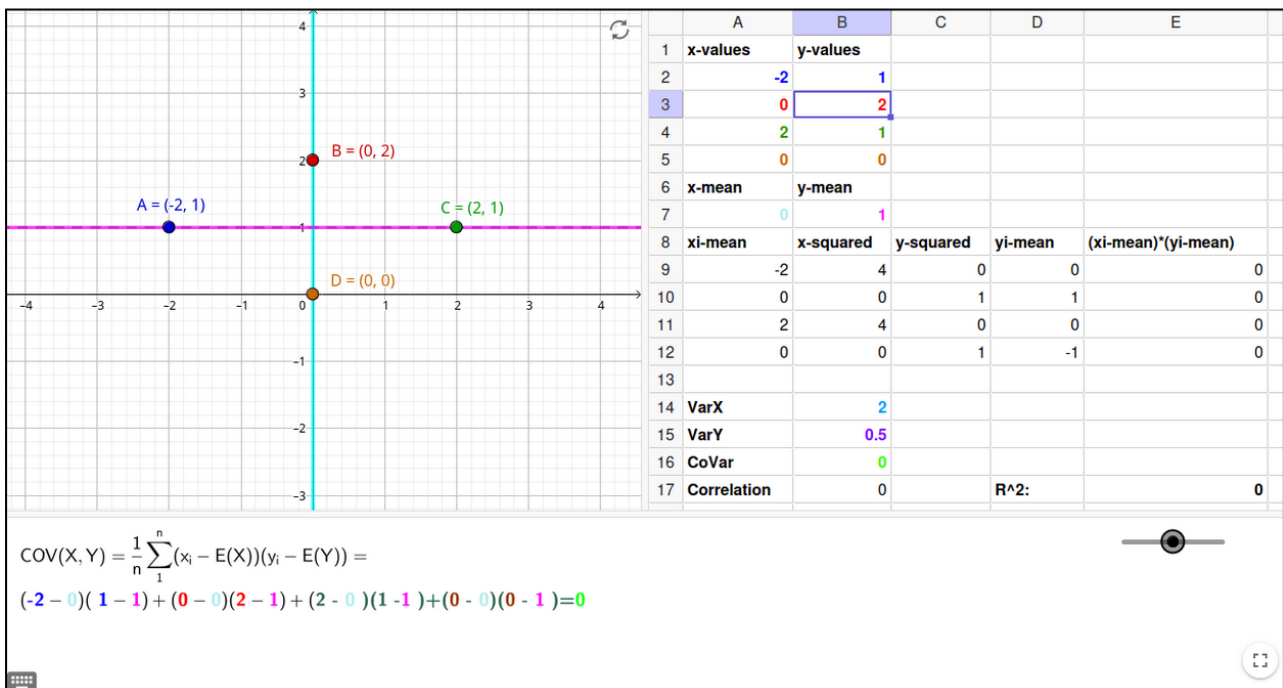
We have noted earlier that when points are, as in this case, clustered close to the y -mean (i.e., the points are “very horizontal”), then $\text{Var}(Y)$ and $\text{Cov}(X, Y)$ are very small. Indeed, in this case,

$\text{Var}(Y) = 0.01$ and $\text{Cov}(X, Y) = 0.02$. Therefore, the arithmetic effect on the formula is that the numerator (i.e., the square of the covariance) is vanishingly small, making the value of R^2 smaller than one might expect. Remember, however, that R^2 is not measuring how well the points fit a line; it is measuring “the proportion of the variance in the response variable that can be explained by the predictor variable in the regression model” (Statology, 2023). In other words, it measures how much the variance in x can explain the variance in y . In this case, where the points are “very horizontal,” there is almost no variance in y so the variance in y is not explained by the variance in x .

Sensitivity of R^2 . It should also be noted that with such small values, R^2 is extremely sensitive to small changes. For example, moving the point at $(1, 0)$ to $(1, 0.2)$ results in an almost 400% increase in R^2 to 0.34. Arithmetically, this can be accounted for by the fact that while the values in the numerator (i.e., $\text{Cov}(X, Y) \cdot \text{Cov}(X, Y)$) are very small, so is at least one value in the denominator—namely, $\text{Var}(Y)$. It’s also worth noting that correlation is undefined when the data points are arranged perfectly horizontally (indicated with a “?”). Many students confuse values that are “undefined” with those that equal 0. *R² Explorer* can help students grasp the distinction between “zero correlation” and “undefined correlation.”

For instance, we encourage our students to drag data points around until they find a configuration of points with a correlation equal to 0 (rather than “?”). *Can you find such a configuration?* Take a few minutes and use the applet to find one or more examples, then read on. Continue by considering the points plotted in Figure 15.

Figure 15. An arrangement of points with correlation equal to 0.

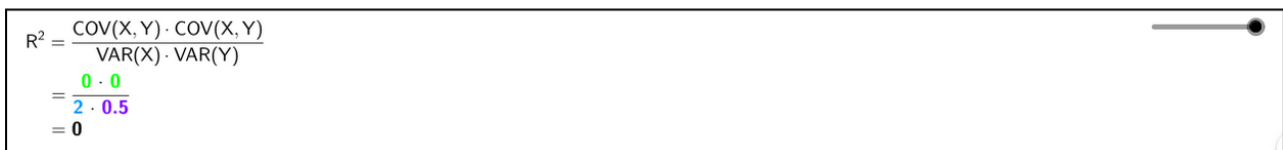


Manipulate your sketch so that we’re graphing the same points, namely $A(-2, 1)$, $B(0, 2)$, $C(2, 1)$, and $D(0, 0)$. Then, drag the slider in the Equation Pane to reveal the formula for $\text{Cov}(X, Y)$, the covariance of x and y , as shown in Figure 15. Note that each plotted point lies on either the solid vertical

or horizontal line. Since the solid **cyan** (vertical) line represents the mean of the x -coordinates, the two points that lie on this line will have x -coordinates equal to the mean, so $x_i - E(X)$ will equal 0 for each. Similarly, $y_i - E(Y)$ will equal 0 for each of the two points that lie on the **magenta** (horizontal) line. So, each partial sum in the $\text{Cov}(X, Y)$ formula will equal 0, and $\text{Cov}(X, Y)$ equals 0.

Dragging the slider in the Equation Pane one position to the right reveals the formula for R^2 for the same data points.

Figure 16. Calculation of R^2 for points A(-2,1), B(0,2), C(-2,1), and (0,0) as depicted in the Equation Pane.



$$R^2 = \frac{\text{COV}(X, Y) \cdot \text{COV}(X, Y)}{\text{VAR}(X) \cdot \text{VAR}(Y)}$$

$$= \frac{0 \cdot 0}{2 \cdot 0.5}$$

$$= 0$$

Since the points aren't arranged horizontally or vertically, variance exists in both x and y , and the covariance will be defined. Note that $\text{Cov}(X, Y)$ is the numerator for covariance. Our previous observations about $\text{Cov}(X, Y)$ help us understand why data points arranged in this manner have R^2 equal to 0.

4 IN CONCLUSION

In summary, the GeoGebra applet provides an interactive and visual platform for students to explore linear regression concepts. Students can develop a more profound conceptual understanding of the statistical concepts involved by actively manipulating data points, observing real-time changes in calculations, and connecting their actions to algebraic representations. The applet's functionality encourages engagement, collaboration, and exploration, allowing students to bridge the gap between graphical observations and mathematical formulas.

With the technology, we can compute complex calculations within seconds, explore many situations, and test different conjectures. However, if we do not discuss or question why these conclusions might be happening, this whole work might stay experimental, and technology might become a black box. Therefore, in this paper, we wanted to unpack the coefficient of determination formula by unpacking other formulas used in these formulas with a step-by-step nature. The coefficient of determination formula seems very simple initially but depends on other formulas. Unpacking these formulas using multiple linked representations allows students to interpret algebraic formulas with the power of visual/graphical and numerical/tabular representations.

REFERENCES

Edwards, M., Meagher, M., and Özgün-Koca, S. A. (2008). Technology tips: When is a good fit not a good fit? Dynamic regression with the TI-Nspire Graphing Calculator. *The Mathematics Teacher*, 102(4):300–305.

NCTM (2000). Principles and standards. <https://www.nctm.org>. Accessed: 2023-07-17.

Statology (2023). R vs. R-Squared: What's the difference? <https://www.statology.org/r-vs-r-squared/>. Accessed: 2023-07-17.



Michael S. Meagher is an Associate Professor in the Department of Secondary Education at Brooklyn College. Dr. Meagher is a former teacher of mathematics and theory of knowledge in Vienna International School. His research interests include the use of technology in the teaching and learning in school mathematics.



Michael Todd Edwards is a professor of mathematics education in the Department of Teaching, Curriculum, and Educational Inquiry (TCE) at Miami University in Oxford, Ohio. His research focuses on the development of technology-oriented curricula, specifically teaching materials that utilize dynamic geometry software and Large Language Models.



S. Asli Özgün-Koca is a professor of mathematics education in the Teacher Education Division at Wayne State University in Detroit, Michigan. She teaches secondary mathematics methods and mathematics education graduate courses in the College of Education. Her research interests include the use of technology in mathematics education and the education and professional development of mathematics teachers.