# DATA DETECTIVE

Author: Kelly Findley and Mary Burr

## PRE-PLANNING

This set of lessons focuses on teaching students to model perfect linear relationships involving bivariate data with an equation and a graph, to use scatterplots and best fit lines to represent imperfect linear relationships, and to distinguish perfect vs. imperfect linear relationships. Students will learn to recognize different kinds of association. Students will also begin to evaluate model fit for imperfect linear relationships by considering how tight a fit is and the range of possible values for one variable given a known value for the other variable.

### LEARNING GOALS

- Students will learn to recognize positive and negative association and reason about these ideas in terms of different real-world contexts
- Students will learn to recognize outliers and clustering
- Students will learn to evaluate whether a relationship is linear or nonlinear
- Students will learn to informally fit a line to bivariate data using only the scatterplot
- Students will learn to define the slope and intercept of the line of best fit and use that as a means of prediction
- Students will consider a range of possible values for one variable given a known value for the other variable and consider tightness and looseness of fit.

### STANDARDS ADDRESSED

- MAFS.8.F.2.4: Construct a function to model a linear relationship between two quantities. Determine the rate of change and initial value of the function from a description of a relationship or from two (x,y) values, including reading these from a table or from a graph. Interpret the rate of change and initial value of a linear function in terms of the situation it models, and in terms of its graph or a table of values.
- MAFS.8.SP.1.1: Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.
- MAFS.8.SP.1.2: Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.
- MAFS.8.SP.1.3: Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.
- Standards of Mathematical Practice
    - M1: Make sense of problems and persevere in solving them

    - M3: Construct viable arguments and critique the reasoning of others
    - M4: Model with mathematics
- Standards of Scientific Practice (we encourage teachers to also consider these practices when teaching statistics!)
    - S4: Analyzing and interpreting data
    - S7: Engaging in argument from evidence
    - S8: Obtaining, evaluating, and communicating information

## CURRICULUM ALIGNMENT

GoMath Grade 8, Modules 5.1 – 5.3 and 14.1 – 14.2

## PRIOR KNOWLEDGE

- Familiarity with the equation of a line and Cartesian coordinate planes.
- Ability to calculate rate using distance and time.

## MATERIALS

- Technology: 2:1 or 1:1 laptop, chromebook, or iPad
- PhET sim: Least-Squares Regression
- Activity sheet
- Florida GoMath: Pre-Algebra (or other curriculum/resources)

# UNIT OUTLINE (5 DAYS)

## LESSONS 5.1 – 5.2

**1-2**
DAYS

Spend the first two days covering topics in lessons 5.1 and 5.2 from the GoMath PreAlgebra Book. The topics included are as follows:

- Writing the equation of a line (slope-intercept form) to model the relationship of two variables within a contextual situation.
- Writing the equation of a line from a graph
- Writing the equation of a line from a table of values.

One optional route for completing these goals are as follows:

- Complete the "Explore Activity" on p. 127
- Complete Example 1 on p. 128
- Complete Example 1 on p. 133
- Complete Example 2 on p. 134
- Also emphasize interpreting slope and intercept in context to the problems (e.g., For y=15+3x, this means that Greta pays $15 a month when she spends 0 hours using the equipment and pays an extra $3 for each hour she uses the equipment)

(Note that students will continue to encounter these types of problems in the next lesson)

## DAY 1

**7**
MINUTES

Warm-up and warm-up discussion (suggestion): "Zara made an initial deposit to a bank account and then added a fixed amount every week. The table shows the money in her account after

| Number of weeks (x) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Balance (y) | $140 | $160 | $180 | $200 | $220 |

a) Write an equation in slope-intercept form to express the situation
b) How much money will she have by the 8$^{th}$ week?"

| | |
|---|---|
| **5** MINUTES | Introduce the Mystery (powerpoint optional).<br>• There's been a robbery at the jewelry store<br>• Two witnesses have each called in to share some information<br>• Read the testimony of Witness A<br>• Read the testimony of Witness B<br>• Present the "Initial Facts" of the Crime |
| **8** MINUTES | Pass out DAY 1 activity sheets and encourage students to "Think-Ink-Pair-Share" about whether they think the witnesses evidence is trustworthy and how we might test their claims against the initial facts.<br>• Provide students a minute to think, a minute to write, and a minute or two to discuss with their neighbor. Use this as an opportunity to promote productive struggle, sharing clarification comments only<br>• Bring class together for discussion. Students may share ideas like, "thief has different sized feet," or "one is a men's size and one is a women's size" (clarify that both are measured using same scale), or connections to F and C temperature or distance/time. Highlight student contributions that seek to link variables. |
| **13** MINUTES | Invite students to work on the temperature section (#1-2) on the second page.<br>• Students who finish early can try the bonus question<br>• Encourage students to think of multiple ways to solve in addition to using the equation (e.g., looking at the graph<br>Recap findings from temperature work and decide if Witness A seems trustworthy**.** Ask students questions like,<br>• "Does Witness A seems trustworthy? Why?<br>• If Witness A is not trustworthy, what might we have expected to see?" |
| **10** | Have students work on next part (#3-6) for distance/time position of thief. |

| | |
|---|---|
| MINUTES | • Again, there are opportunities to use multiple methods, such as comparing the unit rates, judging from the graph, or judging straight from the table |
| **7** MINUTES | Facilitate class-wide discussion about Witness B's claims, and encourage students to engage in claim-evidence reasoning to write final conclusion on page 1.<br><br>*For example: I think Witness A is a reliable source because 25 degrees Celsius is the same as 77 degrees Fahrenheit. Witness B is not trustworthy because the thief would be farther away after 70 seconds than his house.* |
| **5** MINUTES | **Final Report**<br><br>Allow students remaining time to write out a data-based argument (claim and evidence) for the police department. |

**DAY 2**

| | |
|---|---|
| **8** MINUTES | Have students open the sim and play around for a few minutes.<br><br>• As they are writing down things they are familiar and unfamiliar with, take note of what might be new and discussed vs. what can be mentioned in passing.<br><br>Pass out the DAY 2 activity sheet while students explore.<br><br>After 5 minutes of exploring, ask students to volunteer things in the sim that look familiar (likely the scatterplot and the equation of a line) and unfamiliar (likely residuals, correlation coefficient).<br><br>If time,<br><br>• explain that correlation coefficient measures whether there is a strong linear relationship between the two variables *(see teacher appendix at the end of the document for some helpful background!).*<br>• touch on “residuals” and/or “squared residuals” options. Residuals measure the distance between data points and the line of best fit, and squared residuals are another “advanced” way of showing how well a line fits a set of data *(see teacher appendix at the end of the document for some helpful* |

| | |
|---|---|
| | *background!).* |
| **13** MINUTES | **Different Kinds of Trends**<br><br>Demonstrate one of the datasets (e.g. height/shoe size dataset). Ask students what they think the specific dots represent (*subjects or units of measurement we have gathered information from*). Try prefacing by asking how many participants are included in this dataset.<br><br>Have students explore the datasets in the drop-down list and identify relationships that resemble the different scenarios listed to answer #1-6.<br><br>Go over findings from students' data exploration, keeping in mind that there are multiple examples of each and examples that could fit multiple scenarios. Ask students to share aloud their responses to #1-6.<br><br>• Discuss the idea of a strong or weak fit in addition to the scenarios discussed<br>• Make a careful distinction between a strong nonlinear fit (e.g. orbital speed vs distance from the sun) and a weak linear fit (e.g. temperature vs latitude). A relationship can be considered linear, even if it's weak.<br>• After collecting some examples, ask students to consider some additional examples and consider whether the example would be a linear or nonlinear relationship, a perfect or an imperfect linear relationship, a positive or negative association<br>• Examples: Number of apples purchased at store and price paid; number of people completing a job and the number of hours to complete it; distance away from school that a student lives and their grade on a math test |
| **15** MINUTES | **Your Mission**<br><br>• #7: Keep in mind that the idea of "hours with no customers" will be difficult to understand. We suggest explaining to students that the owner likes to know this so that he can monitor how much time clerks had to complete other work like cleaning or filing paperwork |

**Comment [AM1]:** This section could use some better alignment with the activity sheet… I've added in problem numbers where I think the lesson aligns, but if it's unclear for me it is likely to be unclear for teachers.

- OPTIONAL: On the student activity sheets, blank out the variables in #8 and have students decide which axis each variable should be on.

After introducing the task, *have students work in pairs* on #7-8, and move around the classroom to help groups who might feel confused. Do you think this will be a linear or nonlinear relationship? Perfect or imperfect relationship? Positive or negative association?

Have students switch screens with their partner and answer #9 Have them primarily look for differences among the scenarios from the first page (linear or nonlinear, positive, negative, or no association, etc.)

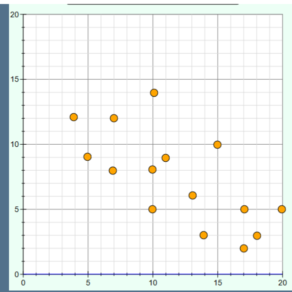
- #10: Have students fit their own line first, and optionally try the "best fit" line *afterwards*. Suggest to students that they have a roughly equal number of points below and above their lines, ignoring outliers. Students should write the equation of the line that appears in the sim.
- #11: Students might need help to know that they should substitute 13 in for x in the equation. Share guiding statements like, "13 hours, is that a possible value for x or for y?"
- #12: you might point students to the fact that the data does not all lie perfectly on the line, so we should also consider the range of possible "Rings stolen" values given 13 hours with no customers, not just our best singular estimate.

**7**

MINUTES

**Complete the Final Report**

*Have a sample scatterplot ready to go to represent what the last 15 weeks could have looked like for the previous activity (suggestion of imperfect, linear, negative relationship). Don't fit a line initially.*




After reviewing the situation, ask students:

- How we might predict our best estimate? (listen for suggestions of fitting a line).
- Can we be sure our estimate is correct or not? Are there other plausible values?
- Come up with a reasonable range of plausible values.

## DAY 3

| | |
|---|---|
| **7** MINUTES | Warm-up and warm-up discussion: teacher's discretion based on what students might need. The scatterplot on p. 434 representing eruptions of Old Faithful might provide for interesting context. Possible path would be to display the scatterplot with context and ask: "A) Is this a perfect or imperfect relationship? B) Is this a positive or negative relationship? C) Are there any outliers in this graph? D) If we looked only at interval times between 70 and 100 minutes, do you think there would be a strong linear association?"<br><br>What's interesting about this case is that there are two clusters of data, and neither one by themselves really has an association between the two variables. |
| **13** MINUTES | Present the newest information about the case.<br><br>• Go over the suspects and their basic info<br>• Go over the three pieces of evidence that have been collected<br><br>Pass out activity sheets and encourage students to "Think-Ink-Pair-Share" about whether they think the witnesses evidence is trustworthy and how we might test their claims against the initial |

| | |
|---|---|
| | facts.<br><br>• May need to explain that BMI is a measure of weight in relation to height, so higher BMI means "thicker" and lower BMI means "thinner." Perhaps show BMI chart on this webpage: http://www.healthyandnaturalworld.com/bmi-chart-and-what-it-can-tell-you/<br>• Provide students a minute to think, a minute to write, and a minute or two to discuss with their neighbor. Use this as an opportunity to promote productive struggle, sharing clarification comments only<br><br>Bring class together for discussion, highlighting comments that connect variables as possibly associated<br><br>Discuss some of the relationships we will look at (IQ/break-in time, height/shoe size) and ask if students think these will have positive associations or negative associations. |
| **20** MINUTES | **Examine the Data** #1-7<br><br>Present datasets that show IQ/time to pick lock, BMI/Waist size, and Height/Shoe size and have students work in groups to discuss how they might narrow down who the thief is.<br><br>To mark the range, students may simply make horizontal tick marks connected by a line<br><br>Circulate and push students for claim-evidence reasoning to justify eliminating suspects.<br><br>A little extra explanation might be needed to explain that for the IQ/time variable, the safe company tested the break-in time of several different people and paired it with their IQ. You can use this data in the table to make a scatterplot before deciding which of the suspects might have reasonably broken in with only 16 minutes. |
| **5** MINUTES | Discuss student ideas for each scatterplot and how they might decide who to eliminate from consideration.<br><br>Push students to explain how they narrowed down the suspects using the scatterplots. |

| | |
|---|---|
| **5** MINUTES | **Final Report**<br><br>From the BMI/waist size scatterplot, students should have eliminated Kenny Kass and Molly Mint. From the shoe/height data, students should have eliminated Jerell Jones and Nina Nash, and from the IQ/break-in time data, students should have eliminated Ravi Raines, leaving Shania Snow as the most likely thief.<br><br>If time, have students reflect on what they learned from this activity sequence |

## Appendix for Teachers

**What is a residual?** A residual is the vertical distance from a data point to the best fit line. Every data point has a residual associated with it. The residual is negative when the data point is below the line of best fit, positive when it is above, and 0 when it is exactly on it.

**What is the best fit line?** The best fit line is the line that minimizes the total residual distance present. While we could place a number of lines on the scatterplot that could be a good fit for the data, only one line will be the *best* fit (minimizes the residual distance to the least possible value).

**Why do we square residuals?** Since some residuals are positive and some are negative, we need a way treat negative residuals the same as positive residuals. One option is to minimize the sum of residual absolute values. Another option is to minimize the sum of squared residuals. The reason we use squared residuals is a bit complicated: The short answer is that a sample will underestimate the presence of large residuals. When minimizing the sum of squared residuals, the larger residuals get more weight, thus creating a best fit line that is more responsive to higher residuals.

**What is the correlation coefficient (r)?** The correlation coefficient measures the **strength** of a *linear* relationship. When the linear relationship is tight, r will approach negative 1 or positive 1, and when there is no linear relationship, r approaches 0. Note that r can still be close to 0 when there are other relationships (like quadratic) in the data! A common misconception for students is that the correlation coefficient is stronger when the slope is steeper. However, a tight linear relationship with a barely positive (or negative) slope can still have a correlation coefficient around 1 or -1. r does not reflect how steep the slope is.

*For more background information on the fitting a line and the "least-squares criterion" (minimizing the sum of squared residuals), see https://onlinecourses.science.psu.edu/stat414/node/277*