Acquisition Lesson Planning Form
Key Standards addressed in this Lesson: MM2D2 a, b, c
Time allotted for this Lesson: 5 Hours

Essential Question: LESSON 3 – Linear Regression
How do you determine the line of best fit using linear and quadratic regression for data using various methods such as eyeballing, median-median line, least squares regression, and technology?
Activating Strategies: (Learners Mentally Active)
Session 1:
Give students 5 scatter plots and have them draw the linear regression line for each one by eyeballing.
Session 2:
Have students write a sentence for each letter in the word MEDIAN. They should share their results with a partner. Then call on various students to share their answer for different letters.
Session 3:
Have students read from the Turnpike Learning Task about the third way to determine the line of best fit, the Least squares method of linear regression. They should complete #4.
Acceleration/Previewing: (Key Vocabulary)
Linear regression;   Median-median line method;   Least squares regression
Teaching Strategies: (Collaborative Pairs; Distributed Guided Practice; Distributed Summarizing; Graphic Organizers)
Session 1: (1 Hour) Using the activating strategy, review with students different lines with various slopes and their equations. Have students guess the equation of the line for each scatter plot. Discuss as a whole group. Use the linear regression line graphic organizer (GO#1 & 2) to teach students how to actually calculate the equation of a linear regression line given a set of data. Be sure to also show students how to do this with their graphing calculators. For practice, give students five sets of data for which they will plot the points, draw a best fit line and calculate the equation of the linear regression line. They should check their results on their calculator. Session 2: (2 Hours) Using the graphic organizer (GO#3 - 5), teach students the method of median-median line for finding the equation of the linear regression line. Teachers should use the interactive website: http://mathplotter.lawrenceville.org/mathplotter/mathPage/medline.htm to demonstrate the median-median line method on their data projector. Students should use the same sets of data from session 1 to find the equation of linear regression line using the median-median line method. Have students compare the results from the different methods used in session 1 and 2.

Math 2   Unit 5 Lesson 3   Linear and Quadratic Regression   Page 1

Session 3: (2 Hours)
In pairs, have students complete #5-6 from the learning task. In whole group, discuss answers to question #6.

So that students can once again practice all three methods, have them complete #7-12 in the learning task.

Session 4: (1 Hour)
Using Traveling on the Turnpike Learning Task Part 2, the students will make a connection between piecewise functions and lines of best fit using least squares regression. Through teacher guided instruction, the students and teacher will complete #13-22.

Session 5: (2 Hours)
Quadratic regression…

Distributed Guided Practice/Summarizing Prompts: (Prompts Designed to Initiate Periodic Practice or Summarizing)

What is the correlation coefficient?

Extending/Refining Strategies:

Summarizing Strategies: Learners Summarize & Answer Essential Question

Session 1:

Traveling on the Turnpike Learning Task Part 1 #1-3. Be sure to write the equation of the line beside the graph.

Session 2:

From the Turnpike Learning Task in session 1, students should find the equation of the linear regression line using the median-median line method. Draw this line on the same graph from session 1. Be sure to write the equation of the line beside the graph.
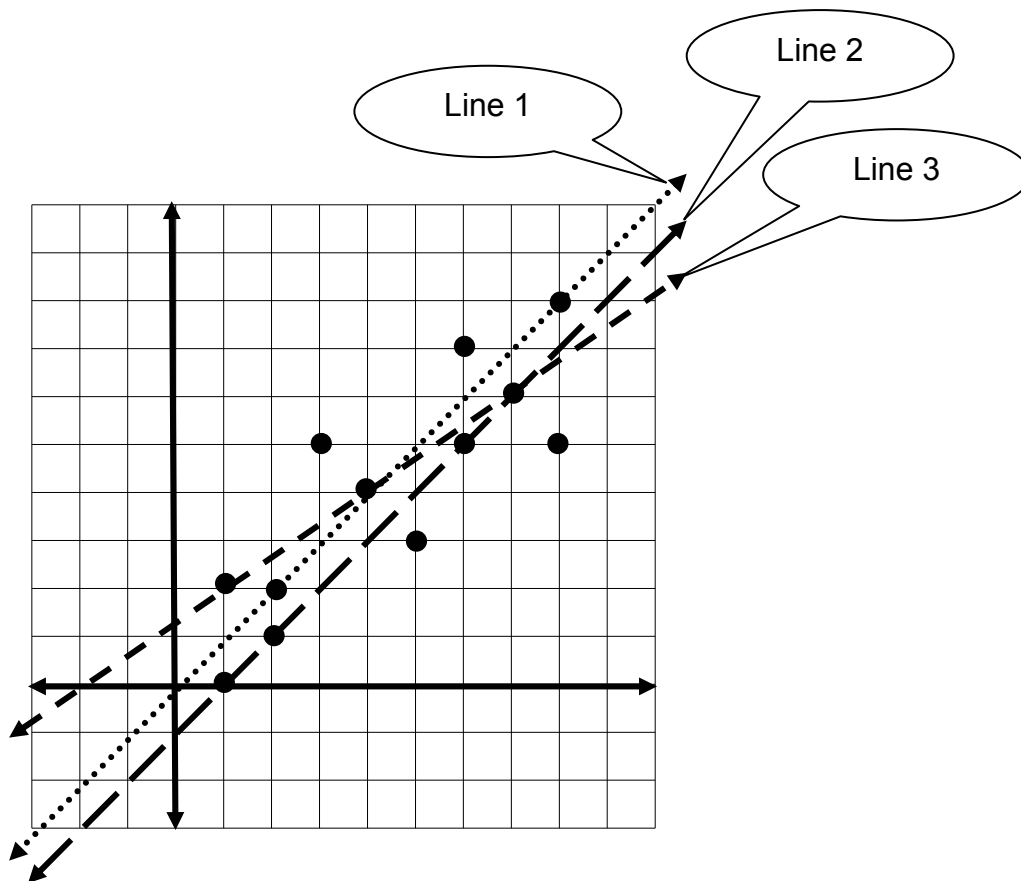
Session 3:

Give students a quiz: Students will be given a set of data of which they should plot the points, eyeball the line of best fit, use visual approximation to write an equation of the line, use the median-median line method to write the equation of the line, and then use technology to find the equation of the line using the least squares regression method. They should compare the three equations and describe which one gives the most accurate representation.

Session 5:

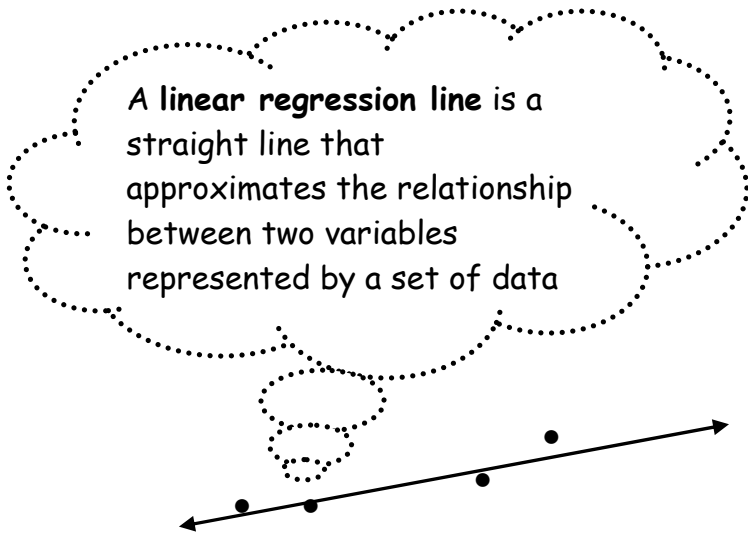# GO #1: How do we eyeball a line best fit a scatter plot?

There are several lines passing through the scatter plot below.
Which one do you think best fits the data?

x	y
1	2
1	0
2	2
2	1
3	5
4	4
5	3
6	5
6	7
7	6
8	5
8	8

Choose two points on each of the three lines and fine the equation of each
before you decide which one fits the best. Show your work below.

GO #2: Linear Regression Lines

A **linear regression line** is a straight line that approximates the relationship between two variables represented by a set of data

Find the linear regression line by calculating the slope

$$\left(m = \frac{y_2 - y_1}{x_2 - x_1} \right) \text{ and using the}$$

slope-intercept form $(y = mx + b)$ of an equation of a line. Check with the linear regression on the graphing calculator.
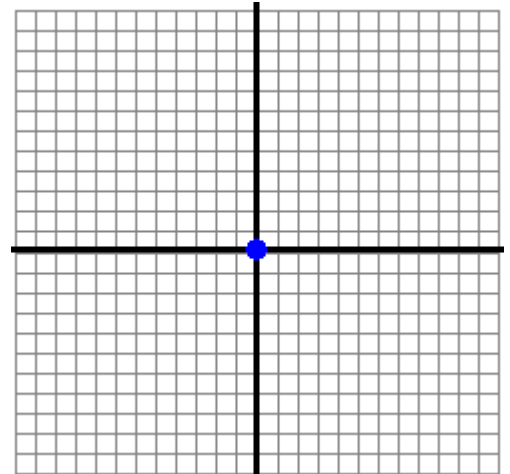
1. Plot the following points on the grid below:
   (1, 3), (2, 2), (0, 1), (3,4)

   By eyeballing:
   Linear Regression Line: $y =$ _____ x + _____

   Using the graphing calculator:
   Linear Regression Line: $y =$ _____ x + _____

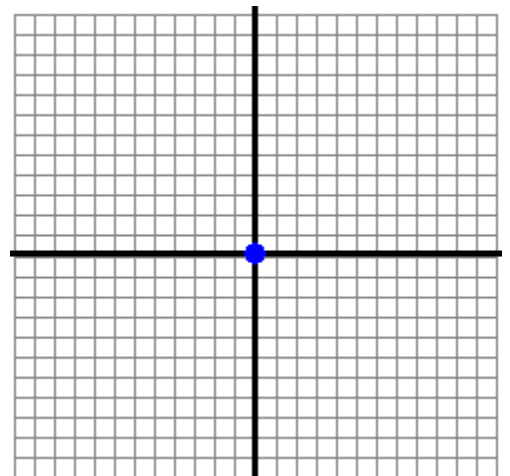2. Plot the following points on the grid below:
   (-4, 2), (1, -2), (-3, -1), (-1, 1)

   By eyeballing:
   Linear Regression Line: $y =$ _____ x + _____

   Using the graphing calculator:
   Linear Regression Line: $y =$ _____ x + _____

GO #3: How do we find the median-median line from a set of data?

1st Divide the points into 3 equal groups. We would have 4 points in each group here. If there were 1 extra point, it would go in the center section. If there were 2 extra points, 1 would be in each of the two outside sections.

| x	y
1	2
1	0
2	2
2	1
3	5
4	2
5	3
6	2
6	7
7	9
8	5
8	8
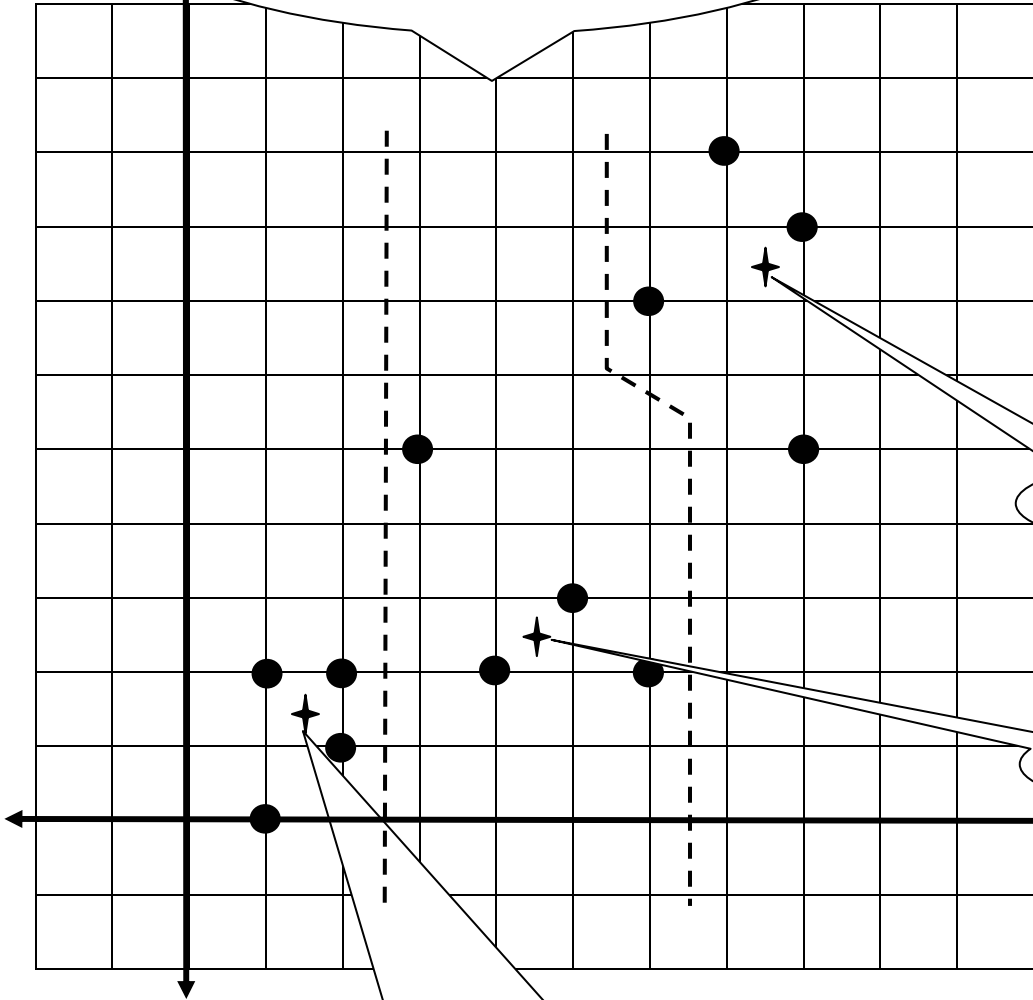
(7.5, 7.5)

(4.5, 2.5)

2nd Find the median x-coordinate and the median y-coordinate in each group of points. This point may or may not be on your graph. For example, in the left section, the median x-coordinate is 1.5 and the median y-coordinate is 1.5, so the point you find is (1.5, 1.5) and is marked by a ✦.

3rd Draw a line through the two points you found in the outside sections. This line may or may not pass through the original points on the graph.

(7.5, 7.5)

(4.5, 2.5)

(1.5, 1.5)

4th Draw a line passing through the point you found in the center section. This line should be drawn parallel to the line you just drew through the outside points. .

5th Draw a line between and parallel to the two lines just drawn. This new line should be 1/3 of the distance from the first line to the second line. In other words, it should be closer to the line through the outside points. This is the **median-median line**.

Now that you have found the median-median line, write its equation. Choose two points on the line and use the point-slope formula to write the equation.
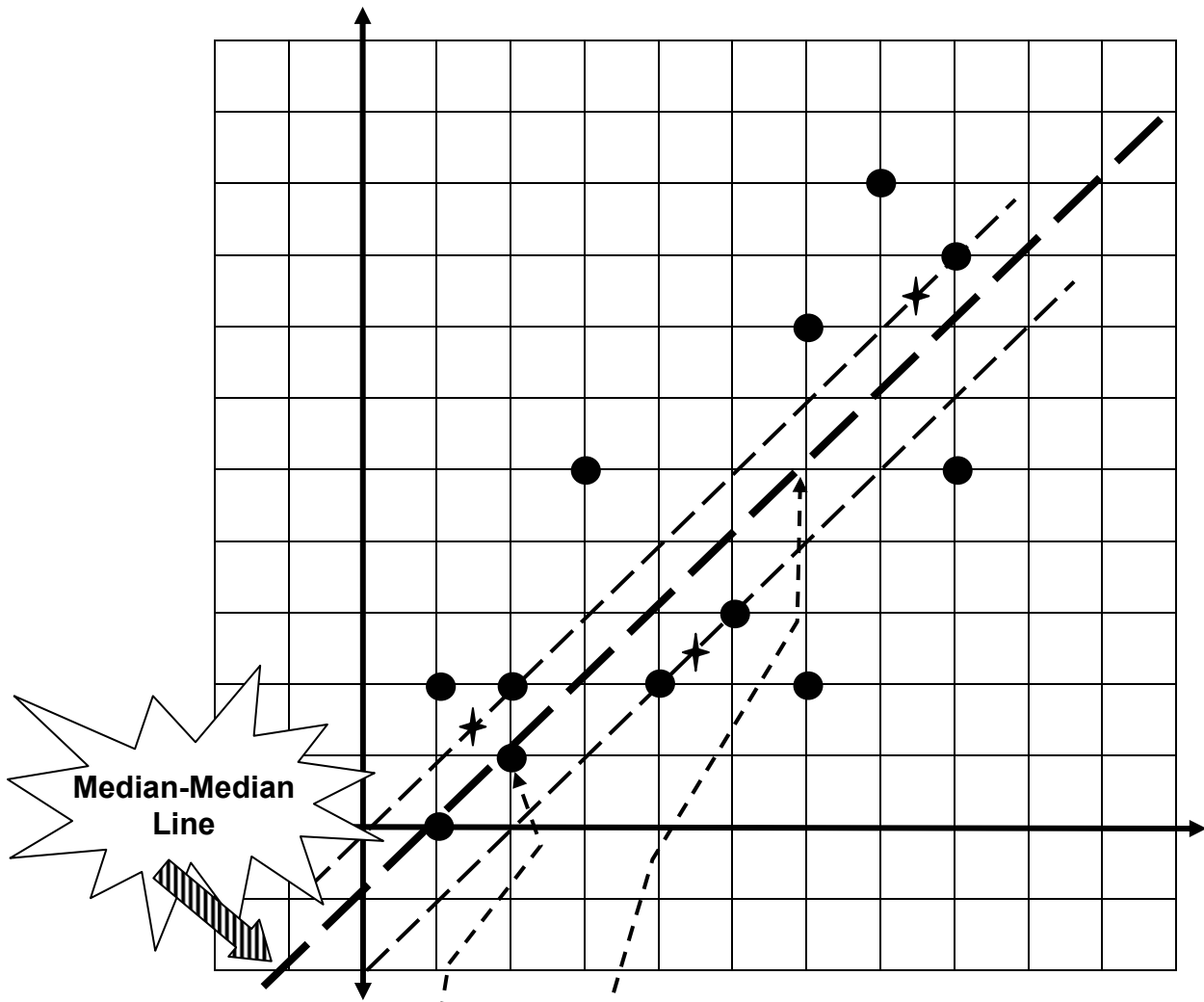
Median-Median Line

Choosing the points (2, 1) and ((6, 5) on the median-median line, the slope would be

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{5-1}{6-2} = \frac{4}{3}.$$

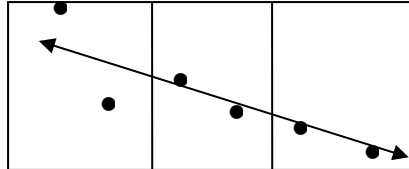Point-slope Formula: $y - y_1 = m(x - x_1)$

$$y - 1 = \frac{4}{3}(x - 2)$$
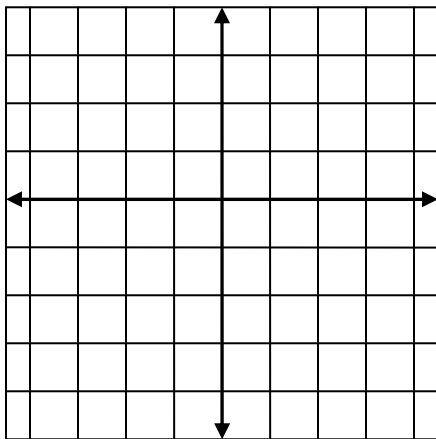
$$y - 1 = \frac{4}{3}x - \frac{8}{3}$$

$y = \frac{4}{3}x - \frac{5}{3}$ is the equation of the median-median line of best fit

GO #4

A **median-median line** is a **linear regression line** that is found by dividing the data into thirds and finding the median x- and y-value of each third. Two lines are drawn using the medians. The regression line lies between these.

1. Plot the following points on the grid below: (-1, -1), (2, 2), (0, 1), (3, 4), (-2, -3), (-3, -4)

2. Divide the data into thirds. Find the three medians.

Median1 = (__,__) Median 2 = (__,__) Median 3 = (__,__)

3. Find the median-median line by hand.

   Median-median Line: $y =$ _____ $x +$ _____

4. Now find the median-median line using the following formula

   Median-median Line Formula $y = ax + b, a = \dfrac{y_3 - y_1}{x_3 - x_1}$, $b = \dfrac{y_1 + y_2 + y_3 - a(x_1 + x_2 + x_3)}{3}$

   Median-median Line: $y =$ _____ $x +$ _____

Are they the same? Why or why not? _____

_____

5. Plot the following points on the grid below: (-1, 0), (0, -1), (2, -1), (-2, 2), (-3, 4), (3, -3)

6. Divide the data into thirds. Find the three medians.

Median1 = (__,__) Median 2 = (__,__) Median 3 = (__,__)

7. Find the median-median line by hand.

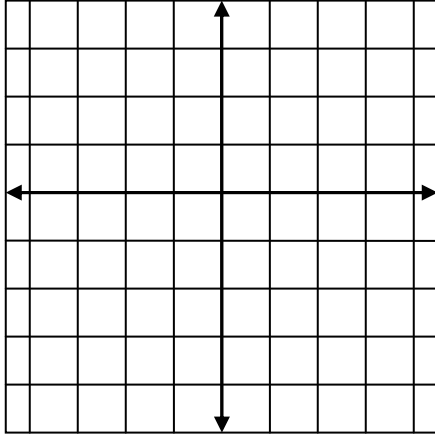   Median-median Line: $y =$ _____ $x +$ _____

8. Find the median-median line using the following formula

Median-median Line Formula $y = ax + b, a = \dfrac{y_3 - y_1}{x_3 - x_1}$, $b = \dfrac{y_1 + y_2 + y_3 - a(x_1 + x_2 + x_3)}{3}$

Median-median Line: $y =$ _____ $x +$ _____

GO #5

A **least squares regression line** minimizes the sum of the squares of the vertical distances between the data points and any possible regression line.

Find the linear regression line by using the graphing calculator.

1. Plot the following points on the grid below: (1, 3), (2, 2), (0, 1), (3,4)
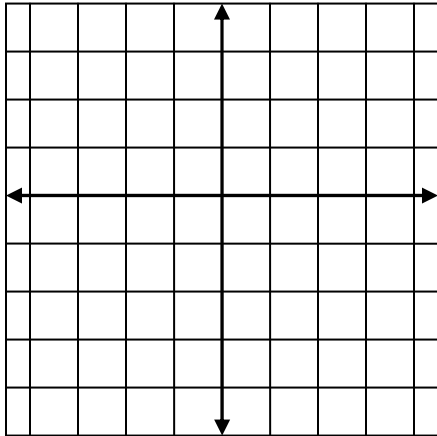
Linear Regression Line: $y =$ _____ $x +$ _____

2. Plot the following points on the grid below: (-4, 2), (1, -3), (0, -1), (-1, 1)

Linear Regression Line: $y =$ _____ $x +$ _____

Traveling on the Turnpike Learning Task – Part I

Examine the toll schedule given by the Ohio Turnpike Commission that is attached to this task. In this task, we will explore the relationship between the distance you travel (based on the numbers of the exits at which you enter and leave the turnpike) and the amount of the toll.

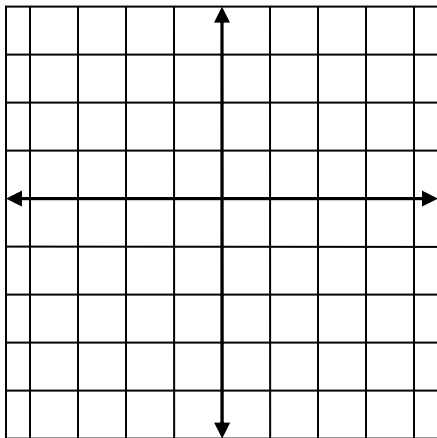1. Suppose you get on the turnpike at Sandusky-Norwalk. Select five exits whose exit numbers are larger than 118 (the Sandusky-Norwalk exit number). For the five exits you select, graph the five ordered pairs of the form (exit #, toll).

2. The points in the data set you selected for *Item 1* are not all on the same line, but the trend of the data could be approximated by a line. Our goal is to write a linear function whose graph provides a "line of best fit" to the points in the scatter plot. Look at the five data points you chose and plotted, and then decide where to draw a line that, in your opinion, best approximates all of the data points. Draw the line on your plot, and explain the reason for your choice of line.

3. Now calculate the equation of the linear regression line based on the graphic organizer in today's lesson.

Data Set 1

Exit numbers larger than exit 118	Toll Fee ($)
135	1.00
140	1.25
142	1.50
145	1.75
151	2.00

Data Set 2

Exit numbers larger than exit 118 through exit 215	Toll Fee ($)
135	1.00
151	2.00
173	3.50
193	4.75
215	6.00
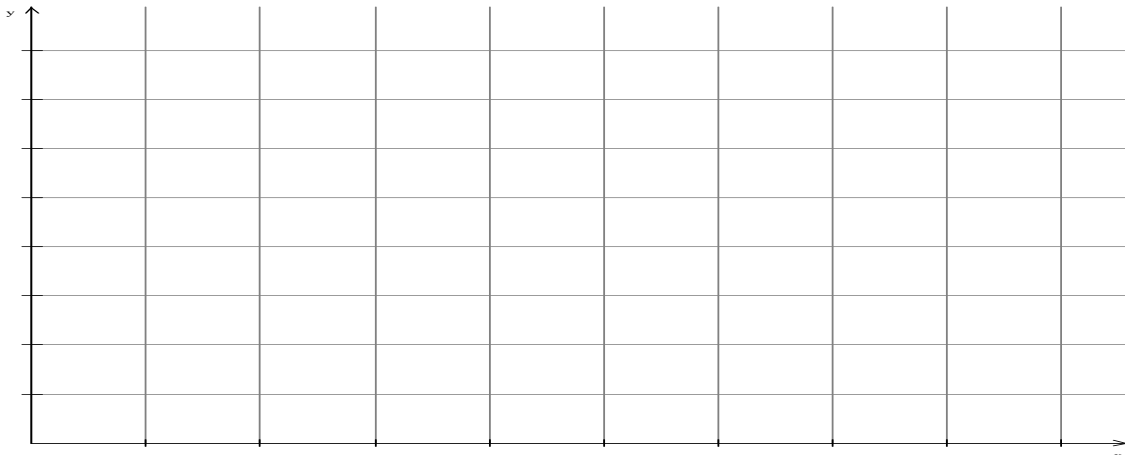
The third way to determine a line of best fit is using the ***least squares method of linear regression***. The least squares method finds the line that minimizes the sum of the squares of the vertical distances from data points to points on the line. We will use technology to find least squares regression lines. The various technological tools that are available use statistical formulas to find the slope and y-intercept of this line. Interested students may want to find out more about these formulas, but our focus in this task is learning characteristics of the least squares regression line and comparing it to visual and median-median regression lines.

4. Use the least squares regression capabilities of your graphing utility to determine a line of best fit for your five data points. Graph this new line on the plot with your data and previous best-fit lines.

5. The least squares regression line is the line that minimizes the sum of the squares of the vertical distances between the data points and the approximation line. All other approximation lines have a greater sum of squares of vertical distances between data points and the approximation line. To verify this property, at least to verify it for the three approximation lines you have created, make three tables using the format indicated below – one for each proposed line of best fit (visual approximation, median-median, least squares regression) – and then compare the sums of the vertical distances and the sums of the squares of the vertical distances.

Exit #	Toll	Predicted toll from the approximation line (using visual approximation)	Vertical distance between point for predicted toll and point for actual toll	Square of the vertical distance between data point and approximation point
		Sum		

Exit #	Toll	Predicted toll from the approximation line (using median-median line)	Vertical distance between point for predicted toll and point for actual toll	Square of the vertical distance between data point and approximation point
		Sum		

Exit #	Toll	Predicted toll from the approximation line (using least squares, Item 4)	Vertical distance between point for predicted toll and point for actual toll	Square of the vertical distance between data point and approximation point
	Sum			

6. The median-median method of linear regression is based on finding the median values of data points. Calculating the least squares method by hand requires using means and standard deviations. From what you already know about medians and means as measures of central tendency, what might be one reason to use the median-median line instead of the more common least squares regression line?

7. Now, let's look at the equation for the tolls if we get on the turnpike at Exit 118 and drove in the other direction. Select five exits whose exit numbers are smaller than 118 (the Sandusky-Norwalk exit number). For the five exits you select, graph the five ordered pairs of the form (exit #, toll).

Data Set 3

Exit numbers less than exit 118	Toll Fee ($)
110	0.50
91	1.75
81	2.25
71	3.00
64	3.25

Data Set 4

Exit numbers less than exit 118 through exit 2	Toll Fee ($)
110	0.50
81	2.25
52	4.00
25	5.75
2	7.25

8. Find a visual regression of line for the five data points you selected in *Item 7*, and draw the line on your scatter plot.
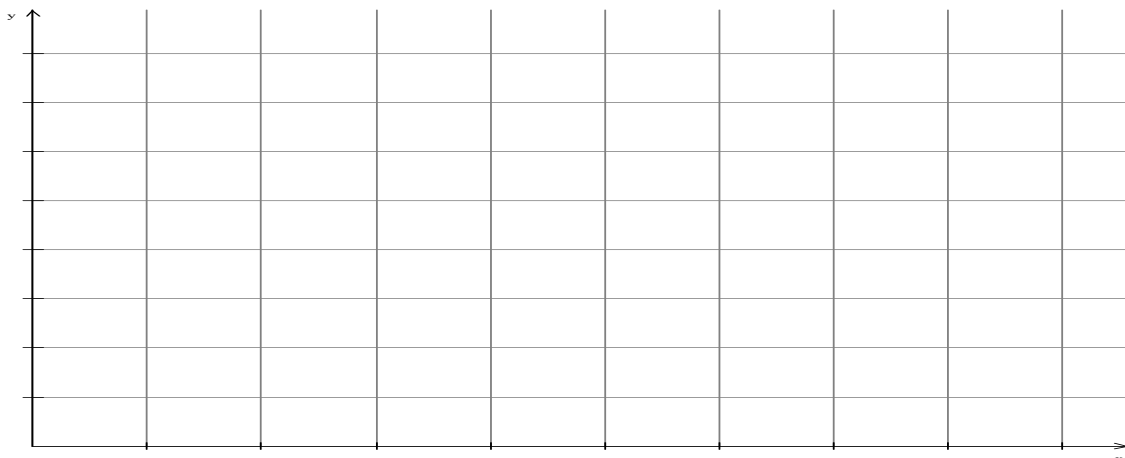
9. Calculat

e the median-median line for the five data points you selected in *Item 7* and then check your results using technology. Draw this line on the same graph with your data points and the line you found by visual linear regression.

10. Using appropriate technology, determine the least squares regression line for the five data points you selected in *Item 7* and draw it on the plot with your data and previous best-fit lines.

11. Consider your least squares regression lines from *Items 4 and 10*. For each line, round the slope to three decimal places and interpret the rounded value in the context of this problem. Explain how you arrive at your interpretation.

12. Consider your least squares regression lines from *Item 10* again. Does the y-intercept have a contextual meaning? If so, what does it represent?

Traveling on the Turnpike Learning Task – Part 2

Our next goal is to find a function that models the toll schedule whenever we begin at the Sandusky-Norwalk exit and travel in either direction. For this part we use two least squares regression lines, one calculated for the points corresponding to all the exits from 118 to 215 and the other calculated for the points corresponding to all the exits from 2 to 118.

13. The least squares regression line for the data set of all ordered pairs of the form (exit #, toll) for the exits from 118 to 215 is given by

$$y = 0.0629011x - 7.46324.$$

    a. Use technology to graph this least squares regression line.

    b. Find the *x*-intercept of the line.

14. The least squares regression line for the data set of all ordered pairs of the form (exit #, toll) for the exits from 2 to 118 is given by

$$y = -0.0621573x + 7.34057.$$

    a. Use technology to graph this least squares regression line.

    b. Find the *x*-intercept of the line.

15. We want to find a function that models the toll schedule whenever we begin at the Sandusky-Norwalk exit and travel in either direction.

    a. How should the graphs of the linear functions from *Items 13 and 14* be restricted so that the graphs can be combined to form a single piecewise function *T* that models the toll schedule for Ohio Turnpike users who start at the Sandusky-Norwalk exit?

    b. Graph the piecewise function *T* from *part a* in a viewing window with $-25 \le x \le 225$ and $-10 \le y \le 10$. If possible, use a graphing utility that allows you to graph piecewise functions. Otherwise, graph by hand to include the required graphing window.

    c. Write the definition of the piecewise function *T* graphed in *part b*.

16. The graph above for *Item 15* may remind you of the graph of the absolute value function, which you first encountered in Unit 1 of Mathematics I. The absolute value function is a piecewise function, and its definition can be written in a piecewise format.

a. Draw the graph of $f(x)=|x|$ using a viewing window with $-5 \le x \le 5$.

b. Identify the pieces of the function *f* that could be defined using a linear function, and then list each linear function and the domain restriction needed to make a part of the graph of *f*.

Function	Domain

c. Combine the pieces listed in *part b* to write the absolute value function as a piecewise function.

17. In your previous work with functions, you have explored the effects of transformations on the graphs of functions. This item asks you to recall some of this information.

a. Explain how the function *g* graphed at the right can be obtained by stretches and shifts of the absolute value function, $f(x)=|x|$.

b. Write a formula for *g(x)* as a transformation of the absolute value function.

c. Write a piecewise definition for the function *g*.

18. We noted that the graph of the function *T* from *Item 15* is similar to the graph of the absolute value function. We can approximate *T* using transformations of the absolute value

function. We'll name this approximation of *T* by a translation of the absolute value function to be the function *S*.

a. Where should the vertex of the graph of the function *S* be located? Why?

b. One of the characteristics of the absolute value functions is its symmetry. What kind of symmetry does the absolute value function have and what symmetry should the graph of *S* have? Why?

c. Consider both parts of the toll graph *T*. Determine a vertical stretch for the absolute value function that you think will produce a function *S* to best approximate *T*. Explain your choice.

d. Based on your answers to *parts a – c* of this item, write a formula for a function *S* that is obtained by transformation of the absolute value function and, in your opinion, best approximates the toll function *T*.

e. How well do you think the function *S* models the actual data from the Ohio Turnpike schedule of tolls and why?

For the remainder of this task, we will focus on the function *S* and use it to analyze additional information about tolls while learning more about absolute value functions.

19. Describe how you would use your function *S* from *Item 18* to determine the amount of the toll if a new exit were added between two existing exits.   Choose a location to add an exit between two existing exits and test your strategy. Where did you place the exit, and how much was the toll?[1] Explain if your prediction is reasonable.

20. We notice in the toll schedule that exits 151 and 152 have the same toll, $2.00. Now that we have a process for finding the toll at future new exits, we could use the underlying idea

---

[1] Question adapted from the NCTM Illuminations lesson "Taking Its Toll." The entire activity can be found at http://illuminations.nctm.org/LessonDetail.aspx?ID=L571 .

to ask new questions such as, "What are all the possible exits that would have a toll of exactly $2.00 if we used the function S to find the initial toll approximation?"

a. Based on the process from *Item 19,* for what toll values predicted by the function S would we set the toll at $2.00?

b. Use the graph of the function S to find all the possible exit numbers for which the process of *Item 19* would yield a toll of exactly $2.00? Explain your process.

21. Using the graph to determine the possible exits that correspond to a particular toll predicted by the function S depends on reading the graph with accuracy. In some cases it may be very difficult to decide if a particular point on the graph of the function S is below, on , or above the line corresponding to a particular y-value line just by examining the graph. For this reason, we need an algebraic method to find the x-value of any points with a particular y-value when the function is an absolute value function. We'll explore the method with a simpler absolute value function and then return to our function S.

a. uppose that we want to find all points on the graph of the function $h(x) = 6|x-3| - 7$ such that $h(x) = 8$. What equation do we need to solve?

b. How can we write the formula for the function h as a piecewise function?

c. Use the piecewise rule for h to replace the one equation given as the answer for *part a* with two different equations to solve, and then solve the two equations.

d. Verify that each of the solutions in part c gives a value of x such that $h(x) = 8$.

22. Now we ask "For what exits, current or future, is the toll on the Ohio Turnpike $5.00 or greater?"

a. Use an algebraic method to answer the question, using the function *S* to find a predicted value of the toll and then rounding as in *Item 19*.

b. Do these solutions agree with the toll schedule?

c. Do these solutions agree with the graph?

OHIO TURNPIKE COMMISSION
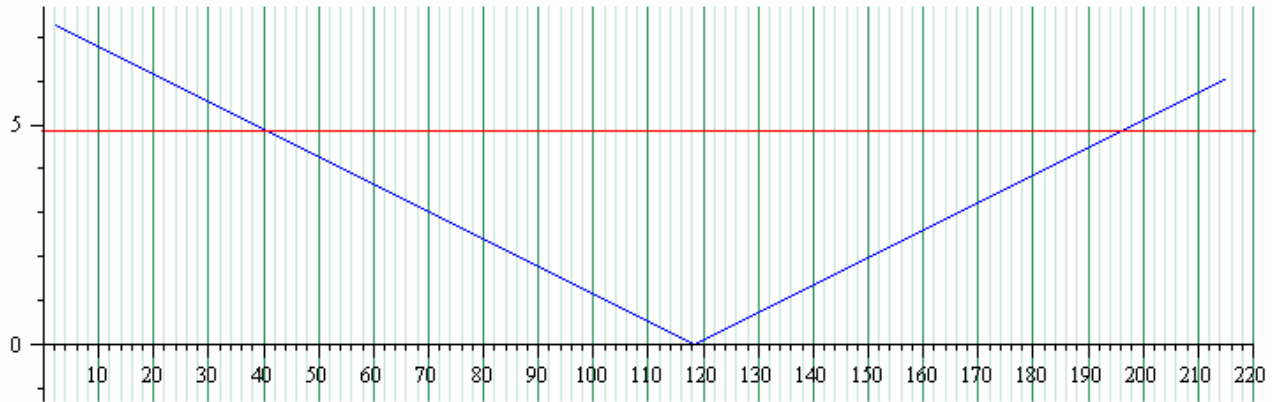
Effective January 1, 2007

SCHEDULE of TOLLS
CLASS 2

Interchange	MP	2	13	25	34	39	52	59	64	71	81	91	110	118	135	140	142	145	151	152	161	173	180	187	193	209	215
Westgate	2		0.75	1.50	2.25	2.50	3.25	3.75	4.00	4.50	5.00	5.75	6.75	7.25	8.50	8.75	9.00	9.00	9.50	9.50	10.00	10.75	11.25	11.75	12.00	13.00	13.25
Bryan-Montpelier	13	0.75		0.75	1.25	1.75	2.50	2.75	3.25	3.50	4.25	4.75	6.00	6.50	7.50	8.00	8.00	8.25	8.50	8.75	9.25	10.00	10.25	10.75	11.25	12.25	12.50
Archbold-Fayette	25	1.50	0.75		0.50	1.00	1.75	2.00	2.50	2.75	3.50	4.00	5.25	5.75	6.75	7.25	7.25	7.50	7.75	8.50	8.50	9.25	9.50	10.00	10.50	11.50	11.75
Wauseon	34	2.25	1.25	0.50		0.50	1.00	1.50	1.75	2.25	3.00	3.50	4.75	5.25	6.25	6.50	6.75	6.75	7.00	7.25	8.00	8.50	9.00	9.50	10.00	10.75	11.25
Delta-Lyons	39	2.50	1.75	1.00	0.50		0.75	1.00	1.25	1.50	2.00	2.50	3.75	4.00	5.00	5.25	5.50	5.50	5.75	6.25	6.25	7.00	7.00	7.50	7.50	9.00	10.00
Toledo Airport-Swanton	52	3.25	2.50	1.75	1.00	0.75		0.50	0.75	1.00	1.50	2.00	3.25	3.50	4.50	4.75	5.00	5.00	5.25	5.50	6.25	7.00	7.50	8.00	8.25	9.25	9.75
Maumee-Toledo	59	3.75	2.75	2.00	1.50	1.00	0.50		0.50	0.75	1.25	1.75	3.00	3.25	4.25	4.50	4.75	4.75	5.00	5.25	6.00	6.75	7.00	7.50	8.00	9.00	9.25
Perrysburg-Toledo	64	4.00	3.25	2.50	1.75	1.25	0.75	0.50		0.50	1.00	1.50	2.75	3.00	4.00	4.25	4.50	4.50	4.75	5.00	5.75	6.50	6.75	7.25	7.50	9.00	9.25
Stony Ridge-Toledo	71	4.50	3.50	2.75	2.25	1.50	1.00	0.75	0.50		0.75	1.25	2.50	2.75	3.75	4.00	4.25	4.25	4.50	4.75	5.50	6.25	6.50	7.00	7.50	8.00	9.00
Elmore-Woodville-Gibsonburg	81	5.00	4.25	3.50	3.00	2.00	1.50	1.25	1.00	0.75		0.50	1.75	2.25	3.25	3.75	3.75	3.75	4.25	4.00	4.25	5.00	5.25	5.75	6.25	7.75	8.25
Fremont-Port Clinton	91	5.75	4.75	4.00	3.50	2.50	2.00	1.75	1.50	1.25	0.50		1.00	1.50	2.75	2.75	3.25	3.25	3.75	3.75	4.25	5.00	5.25	5.75	6.25	7.25	7.75
Sandusky-Bellevue	110	6.75	6.00	5.25	4.75	3.75	3.25	3.00	2.75	2.50	1.75	1.00		0.50	1.75	2.00	2.25	2.50	2.50	2.75	3.25	4.00	4.25	4.75	5.25	6.25	6.50
Sandusky-Norwalk	118	7.25	6.50	5.75	5.25	4.00	3.50	3.25	3.00	2.75	2.25	1.50	0.50		1.00	1.25	1.50	1.50	2.00	2.00	2.75	3.50	3.75	4.25	4.75	5.75	6.00
Vermilion	135	8.50	7.50	6.75	6.25	5.00	4.50	4.25	4.00	3.75	3.25	2.75	1.75	1.00		0.50	0.50	0.75	1.00	1.00	1.50	2.25	2.50	3.00	3.50	4.50	5.00
Amherst-Oberlin	140	8.75	8.00	7.25	6.50	5.25	4.75	4.50	4.25	4.00	3.75	2.75	2.00	1.25	0.50		0.50	0.50	0.75	0.75	1.25	2.00	2.50	2.50	3.00	4.00	4.75
Lorain Co. West	142	9.00	8.00	7.25	6.75	5.50	5.00	4.75	4.50	4.25	3.75	3.25	2.25	1.50	0.50	0.50		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lorain-Elyria	145	9.00	8.25	7.50	6.75	5.50	5.00	4.75	4.50	4.25	3.75	3.25	2.50	1.50	0.75	0.50	N/A		0.75	1.00	1.75	2.50	2.75	3.25	3.25	4.25	4.25
N.Ridgeville-Cleveland	151	9.50	8.50	7.75	7.00	5.75	5.25	5.00	4.75	4.50	4.25	3.75	2.50	2.00	1.00	0.75	N/A	0.75		N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
N.Olmsted-Cleveland	152	9.50	8.75	8.50	7.25	6.25	5.50	5.25	5.00	4.75	4.00	3.75	2.75	2.00	1.00	0.75	N/A	1.00	N/A		0.50	1.25	1.50	2.00	2.25	3.50	4.00
Strongsville-Cleveland	161	10.00	9.25	8.50	8.00	6.25	6.25	6.00	5.75	5.50	4.25	4.25	3.25	2.75	1.50	1.25	N/A	1.75	N/A	0.50		0.75	1.25	1.50	2.00	3.00	3.25
Cleveland	173	10.75	10.00	9.25	8.50	7.00	7.00	6.75	6.50	6.25	5.00	5.00	4.00	3.50	2.25	2.00	N/A	2.50	N/A	1.25	0.75		0.50	1.00	1.25	2.25	2.50
Akron	180	11.25	10.25	9.50	9.00	7.00	7.50	7.00	6.75	6.50	5.25	5.25	4.25	3.75	2.50	2.50	N/A	2.75	N/A	1.50	1.25	0.50		0.50	0.75	1.75	2.25
Streetsboro	187	11.75	10.75	10.00	9.50	7.50	8.00	7.50	7.25	7.00	5.75	5.75	4.75	4.25	3.00	2.50	N/A	3.25	N/A	2.00	1.50	1.00	0.50		0.50	1.25	1.75
Ravenna	193	12.00	11.25	10.50	10.00	7.50	8.25	8.00	7.50	7.50	6.25	6.25	5.25	4.75	3.50	3.00	N/A	3.25	N/A	2.25	2.00	1.25	0.75	0.50		1.00	1.25
Warren	209	13.00	12.25	11.50	10.75	9.00	9.25	9.00	9.00	8.00	7.75	7.25	6.25	5.75	4.50	4.00	N/A	4.25	N/A	3.50	3.00	2.25	1.75	1.25	1.00		0.50
Lordstown-West	215	13.25	12.50	11.75	11.25	10.00	9.75	9.25	9.25	9.00	8.25	7.75	6.50	6.00	5.00	4.75	N/A	4.25	N/A	4.00	3.25	2.50	2.25	1.75	1.25	0.50	

Orbital Debris Learning Task

Whenever a space shuttle is launched, mission-related debris are released as part of the mission. Satellites that have exhausted their missions remain in orbit, and the bodies of many of the rockets used to launch various spacecraft are still in space. Paint chips off spacecraft continue to float in orbit long after the spacecraft has returned to earth. All of these man-made items contribute to the debris in orbit about the Earth. The area where orbital debris is most concentrated is known as the LEO (low Earth orbit) and consists of space 200 and 2000 km above the Earth's surface. Telecommunications satellites are found in the geostationary orbit, above the LEO; according to the orbital debris specialists, there is little debris in the geostationary orbit.

Orbital debris can collide with current and future spacecraft and with each other. Although all spacecraft collide with small particles during a mission, a collision with a particle 10 cm in diameter (the measure across the widest part) or larger can cause catastrophic damage to a craft.

In response to the growing space pollution issue of orbital debris, the United States began close monitoring of orbital space debris, and, along with the United Nations, put into effect a plan to reduce the amount of orbital debris produced and left in space each year. In the United States, a plan for minimizing the creation of new orbital debris was proposed in 1997 and approved in 2001. One provision of this plan includes engineering space shuttle missions and satellite launches so that mission-related debris reenters the Earth's atmosphere and either burns up in descent through the atmosphere or falls safely to Earth in an uninhabited area.

All debris with a diameter of 10 cm or greater in low Earth orbit and with a diameter of 1 meter or greater in geostationary Earth orbit (GEO) are catalogued by the United States Space Surveillance Network. There are also hundreds of thousands of uncataloged smaller particles in Earth orbit. The work of monitoring and reporting orbital debris is carried out at the National Aeronautical and Space Administration (NASA) Orbital Debris Program Office at the Johnson Space Center in Houston, Texas. Scientists at the Johnson Space Center use sophisticated mathematical models and complex computer software to track orbiting debris and space collisions involving orbital debris. Tracking debris is a complex activity because the debris count changes so often. Throughout each year, new spacecraft are launched adding rocket bodies and other mission debris, and break-ups and collisions fragment large debris into many smaller pieces. While these processes add to the debris, some debris that has been in orbit for a long time falls back into the atmosphere to return to Earth or be burned up on reentry. In this task, we will model orbital debris data while we explore additional concepts about modeling data.

1. At the beginning of October 1997, there were 8545 man-made objects in Earth orbit. Over seventy percent of these were orbital debris: mission-related discards, rocket bodies, and fragmentation debris from break-ups and collisions. By late September 2006, there were 9800 man-made objects in Earth orbit, about seventy percent of which were debris.

    a. The table below lists the total number of orbital debris, as cataloged by the US Space Surveillance Network, for the years listed in the table[2].

---

[2] Data taken from October issues of the Orbital Debris Quarterly Newsletter available online at http://orbitaldebris.jsc.nasa.gov/newsletter/newsletter.html . According to Nicholas L. Johnson,

Year	1997	1998	1999	2000	2001	2004	2005	2006
Man-made Orbital Debris (number of objects)	6166	6132	5980	6134	6092	6307	6480	6791

The first man-made object to orbit the Earth was Sputnik 1, launched by the Union of Soviet Socialist Republics (USSR) on Oct. 4, 1957. The debris totals provided here are from dates near the end of the third quarter of each year. Thus, the dates for the data represent a whole number of years from the Sputnik 1 launch, and we can simplify our data by using the number of years since the launch as the first coordinate. Write points corresponding to the data in the table using the form
(number of years since Sputnik 1 launch, number of man-made orbital debris).

b. Make a scatterplot of the points from *part a*.

c. Explain any trends in the data.

One of our topics for this task is consideration of methods for deciding what type of function to use to model a set of data, that is, deciding whether to use a linear model for the data or use a model based on another type of function, such as a quadratic function. Note that data that are best modeled by only the increasing, or only the decreasing, part of a quadratic function may appear linearly related .

Statisticians use a number, called the **correlation coefficient,** as a measure of the extent to which the data for two variables show a linear relationship. The correlation coefficient, usually denoted by **r**, measures the both the direction and strength of a *linear* relationship between two variables.

If the relationship between the data for two variables *exactly* fits a linear function with positive slope, the correlation coefficient for the set of data is 1. The number 1 is largest possible value for a correlation coefficient. A positive correlation coefficient indicates that a linear approximation of the data would have a positive slope. Thus, when a set of data has a positive correlation coefficient, an increase in one variable is associated with an increase in the other variable, and a decrease in one variable is generally associated with a decrease in the other variable. The closer a positive correlation coefficient is to 1, the stronger is the linear

relationship between the variables; the closer a positive correlation coefficient is to 0, the weaker the relationship.

If the relationship between the data for two variables *exactly* fits a linear function with a negative slope, the correlation coefficient is –1. The number –1 is smallest possible value for a correlation coefficient.   A negative correlation coefficient indicates that a linear approximation of the data would have a negative slope. Thus, when a set of data has a negative correlation coefficient, an increase in the value of one variable is generally associated with a decrease in the value of the other variable. The closer a negative correlation coefficient is to –1, the stronger the relationship; the closer a negative correlation coefficient is to 0, the weaker the relationship.

A correlation coefficient of zero means that there is no linear relationship for the points in the set of data.

The formula for correlation coefficient is given by the equation: $r = \dfrac{1}{n-1}\sum\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)$,

where (x_1, y_1), (x_1, y_1), ..., (x_1, y_1) are the data points, $\bar{x}$ is the mean of the *x*-values, $\bar{y}$ is the mean of the *y*-values, s_x is the standard deviation of the *x*-values, and s_y is the standard deviation of the *y*-values. In this task, you are expected to use appropriate technology, either a calculator or software package, to calculate the correlation coefficient.

There is one important caution about correlation coefficient that must be emphasized: a strong correlation, even one with a coefficient of 1 or –1, does not indicate a causal relationship between the variables. We will explore this caution further in another task.

2. Use a calculator or other appropriate technology to find the correlation coefficient, *r*, for your data points from *Item 1*.

    a. Based on the value of the correlation coefficient, what are the strength (weak, moderate, or strong) and direction (positive or negative) of any linear relationship for this set of data?

    b. Using appropriate technology (graphing calculator, Excel, TI-Interactive, etc.), determine the least-squares regression line for the data.

    c. Graph the least squares regression line from *part b* together with a scatterplot of the data.

d. How well does the least squares regression line appear to fit the data?

e. Find the sum of the squares of the vertical distances between points on the regression line and the corresponding data points.

f. Interpret the slope of your least squares regression line from *part b*.

g. Interpret the y-intercept of your squares regression line from *part b*.

Many data sets studied by scientists and engineers, like the one studied in Items 1 and 2, can be viewed as functions since each input has a unique output. However, such real-life data are usually not modeled exactly by any function that we can build using basic functions or transformations of basic functions. However, mathematical models built from transformations of basic functions can provide useful information about the data. Thus, finding and comparing different models is a useful endeavor.

Linear functions are not the only functions we can use to model data. In the "Traveling on the Turnpike" task, you used least squares linear regression to build an absolute value function to model that data. In the remainder of this task, we explore using quadratic functions to model data, called ***quadratic regression***. Modeling data with other polynomial functions and exponential functions is also possible. Although our study is limited to quadratic regression, we will see concepts the generalize for higher degree polynomials.

3. This item explores quadratic regression for the data points from *Item 1*. Quadratic regression finds the quadratic function that minimizes the sum of the squares of the vertical distances between points on the quadratic graph and the data points.
    a. Use a calculator or other appropriate technology to find the quadratic regression equation for the data.

    b. Graph the quadratic regression curve from *part b* together with a scatterplot of the data.

c. How well does the parabola appear to fit the data?

d. Find the sum of the squares of the vertical distances between points on the quadratic regression curve and the corresponding data points.

e. Which function, the least squares linear regression or the quadratic regression, provides a better model for the data from *Item 1*? Explain your answer.

f. The data in *Item 1* were obtained from the quarterly news letter published by the NASA Orbital Debris Program Office. The newsletter was not published for a period including the third quarters of 2002 and 2003 so, for these two years, the count of man-made debris in Earth orbit at the end of the third quarter was not available. Use your choice of the better model (from *part e*) to predict the amount of debris for these two years.

Finding the answers for *Item 3, part f* is an example of **interpolation**. Interpolation is used to make predictions within the domain of values of the independent variable. It is standard to find a regression curve for the purpose of finding one or more interpolated values.

Extrapolation is the use of a regression curve to make predictions outside the domain of values of the independent variable. Care must be taken to determine the feasibility of extrapolation; that is, sufficient evidence should exist that the trend described by the regression curve would continue outside the present domain.

4. Consider using your regression models for the data in *Item 1* to estimate man-made orbital debris during the period 1957 – 1966.

a. Discuss whether it is feasible to use your least squares linear regression to extrapolate values for 1957 – 1966.

b. Discuss whether it is feasible to use your quadratic regression function to extrapolate values for 1957–1966.

5. The amount of man-made debris can change in an instant when a satellite breaks apart or there is a collision among two pieces of debris. Approximately 1500 new debris were added in 2007 as a result of the breakup of the Fengyun-1C and Upper Atmospheric Research Satellites; these events made 2007 the worst ever year for producing new man-made debris.

   a. The actual count of man-made debris in Earth orbit at the end of the third quarter of 2007 was 9250. State the error in using your linear and quadratic models to predict the count of man-made debris in Earth orbit at the end of the third quarter of 2007?

   b. The actual count of man-made debris in Earth orbit at the end of the third quarter of 2008 was 9661. State the error in using your linear and quadratic models to predict the count of man-made debris in Earth orbit at the end of the third quarter of 2008?

   c. If you found a new linear or quadratic regression function using the data from *parts a and b* above as well as the data from *Item 1, part a,* would you be confident in using your model above to predict the amount of man-made debris in Earth orbit for 2010? Explain.

6.

1990	10			
1995	19			
2000	28			
2005	40			

a. Based on 4a and your regression curves and plots in problem 2, which curve is a better fit for the data? Explain your answer.

b. Use your choice from 4b to predict the cumulative number of LEO collisions in 2010.

Collisions in the LEO
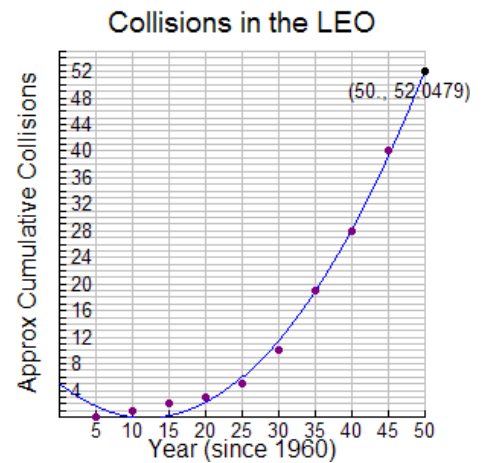
Causation versus Correlation Learning Task

STATS is a non-profit, non partisan research organization affiliated with George Mason University. The following is reproduced from their website (http://www.stats.org/faq_vs.htm

One of the most common errors we find in the press is the confusion between *correlation* and *causation* in scientific and health-related studies. In theory, these are easy to distinguish — an action or occurrence can *cause* another (such as smoking causes lung cancer), or it can *correlate* with another (such as smoking is correlated with alcoholism). If one action causes another, then they are most certainly correlated. But just because two things occur together does not mean that one caused the other, even if it seems to make sense.

Unfortunately, our intuition can lead us astray when it comes to distinguishing between causality and correlation. For example, eating breakfast has long been correlated with success in school for elementary school children. It would be easy to conclude that eating breakfast *causes* students to be better learners. It turns out, however, that those who don't eat breakfast are also more likely to be absent or tardy — and it is absenteeism that is playing a significant role in their poor performance. When researchers retested the breakfast theory, they found that, independent of other factors, breakfast only helps undernourished children perform better.

Many many studies are actually designed to test a correlation, but are suggestive of "reasons" for the correlation. People learn of a study showing that "girls who watch soap operas are more likely to have eating disorders" — a correlation between soap opera watching and eating disorders — but then they incorrectly conclude that watching soap operas *gives* girls eating disorders.

In general, it is extremely difficult to establish causality between two correlated events or observances. In contrast, there are many statistical tools to establish a statistically significant correlation.

There are several reasons why common sense conclusions about cause and effect might be wrong. Correlated occurrences may be due to a common cause. For example, the fact that red hair is correlated with blue eyes stems from a common genetic specification which codes for both. A correlation may also be observed when there is causality behind it — for example, it is well-established that cigarette smoking not only correlates with lung cancer, but actually causes it. But in order to establish cause, we would have to rule out the possibility that smokers are more likely to live in urban areas, where there is more pollution — or any other possible explanation for the observed correlation.

In many cases, it seems obvious that one action causes another. However, there are also many cases when it is not so clear (except perhaps to the already-convinced observer). In the case of soap-opera watching anorexics, we can neither exclude nor embrace the hypothesis that the television is a cause of the problem — additional research would be needed to make a convincing argument for causality. Another hypothesis is that girls inclined to suffer poor body image are drawn to soap operas on television because it satisfies some need related to their poor body image. Yet another hypothesis is that neither causes the other, but rather there is a common trait — say, an overemphasis on appearance by the girls' parents — that causes both an interest in soap operas and an inclination to develop eating disorders. None of these

hypotheses are tested in a study that simply asks who is watching soaps and who is developing eating disorders, and finding a correlation between the two.

How, then, does one ever establish causality? This is one of the most daunting challenges of public health professionals and pharmaceutical companies. The most effective way of doing this is through a *controlled study*. In a controlled study, two groups of people who are comparable in almost every way are given two different sets of experiences (such one group watching soap operas and the other game shows), and the outcome is compared. If the two groups have substantially different outcomes, then the different experiences may have caused the different outcome.

There are obvious ethical limits of controlled studies – it would be problematic to take two comparable groups and make one smoke while denying cigarettes to the other in order to see if cigarette smoking really causes lung cancer. This is why epidemiological (or observational) studies are so important. These are studies in which large groups of people are followed over time, and their behavior and outcome is also observed. In these studies, it is extremely difficult (though sometimes still possible) to tease out cause and effect, versus a mere correlation.

Typically, one can only establish correlation unless the effects are extremely notable *and* there is no reasonable explanation that challenges causality. This is the case with cigarette smoking, for example. At the time that scientists, industry trade groups, activists and individuals were debating whether the observed correlation between heavy cigarette smoking and lung cancer was causal or not, many other hypotheses were considered (such as sleep deprivation or excessive drinking) and each one dismissed as insufficiently describing the data. It is now a widespread belief among scientists and health professionals that smoking does indeed *cause* lung cancer.

When the stakes are high, people are much more likely to jump to causal conclusions. This seems to be doubly true when it comes to public suspicion about chemicals and environmental pollution. There has been a lot of publicity over the purported relationship between autism and vaccinations, for example. As vaccination rates went up across the United States, so did autism. However, this correlation (which has led many to conclude that vaccination causes autism) has been widely dismissed by public health experts. The rise in autism rates is likely to do with increased awareness and diagnosis, or one of many other possible factors that have changed over the past 50 years.

In general, we should all be wary of our own bias; we like explanations. The media often concludes a causal relationship among correlated observances when causality was not even considered by the study itself. Without clear reasons to accept causality, we should only accept correlation. Two events occurring in close proximity does not imply that one caused the other, even if it seems to makes perfect sense.

Do the following activities with the aid of research on the topic of correlation versus causation. An Internet search on this topic will produce a wealth of examples to consider. Most introductory statistics textbooks have a section on the topic.

1. Find one example, or collect data to create an example of your own, of a strong correlation (either positive or negative) between variables where the idea that either one causes the other is clearly not reasonable.

   a) Describe the two variables that are correlated.

b) Give the correlation coefficient or evidence to support the claim that there is a strong correlation.

c) Write a sentence making the claim that one variable causes the second.

d) Give an explanation of why the correlation may exist.

e) Make a bibliography of all sources used.

2. Find an example of a topic of current public interest where there is a correlation between two quantities, call them A and B, that has been interpreted, at least in some media, as A causing B.

a) State the claim that A causes B providing the source of the claim and explanations of terms as necessary.

b) Discuss possible reasons that there may be a correlation without A actually causing B, such as a third variable C related to A and B.

c) Make a bibliography of all sources used.

View Tube Learning Task

This learning task requires you to work in pairs to collect and then analyze data.

Supplies needed by each pair of students:
- a cardboard tube, such as a paper towel or gift roll
- yard stick or meter stick
- masking tape
- marker
- strips of colored paper

Set up for data collection:
- Choose one student to be the viewer and one to be the recorder.
- Choose a wall that is unobstructed for several feet in front of the wall and to which objects can be temporarily affixed with masking tape.
- Use masking tape to create a line on the floor to measure the distance of an observer from the wall and use the yard stick, or meter stick, to set up a scale on the masking tape line.
- Tape the yard stick, or meter stick, to the wall so that it is oriented vertically and is positioned above the point where the masking tape line intersects the wall. The center value of the yard stick, or meter stick, should be at eye level, approximately, for the viewer.

Data collection:
- **Use the following steps for collecting each data.**
  - (i) The viewer chooses marked distance from the wall, stands at this mark, and looks at the wall through the view tube, (being careful to hold it steady)
  - (ii) The recorder used colored strips of paper and oral communication with the viewer to determine the highest point and the lowest point on the yard stick, or meter stick, that are visible through the tube.
  - (iii) The recorder writes down the viewer's distance from the wall and the values of the highest and lowest points visible to the viewer through the view tube.
- Collect data for at least 6 different distances from the wall and increment the distance from the wall by the same amount each time as successive data is collected.
- For each distance from the wall, calculate, the *diameter* of the viewing circle and the *area* of the viewing circle.

1. Consider the relationship between the diameter of the viewing circle and the distance from the wall.

a. Why is this relationship a function?

b. What is the independent variable for the function?

c. Using appropriate technology, graph the function for this relationship.

d. Using appropriate technology (graphing calculator, Excel, TI-Interactive, etc.), determine the least squares regression line and the correlation coefficient for the data.
e. Graph the least squares regression line on the same axes with the data.

f. Describe the direction and strength of the relationship between the diameter of the viewing circle and the distance from the wall.

2. Consider the relationship between the area of the viewing circle and the distance from the wall.
   a. Why is this relationship a function?

   b. What is the independent variable for the function?

   c. Using appropriate technology, graph the function.

   d. Using appropriate technology, determine the least squares quadratic regression for the data.

   e. Graph the least squares quadratic regression on the same axes with the data.

In *Items* 1 and 2, you calculated regression curves to get a line or quadratic of best fit to the data. We now explore a mathematical tool that is used to determining whether a set of data fits a linear or quadratic model; this tool is called the **method of finite differences**.

3. Consider the basic linear equation $y = 2x - 3$. We know that the graph of this equation is a straight line.
   a. What are the slope and y-intercept of this line is 2?

   b. Complete the table of values below. In the last column, determine the difference in each y-value and the one immediately preceding it. (The Greek letter delta, "Δ", which corresponds to the English letter "D", is used to indicate the difference, or change in, a quantity. Thus, the last column is labeled Δy.)

x	y	Δy
−2		
−1		
0		
1		
2		

   c. What do you notice about the successive differences?

   d. How do the successive differences relate to the equation of the line? Why?

   e. Complete the table of values below and, in the last column, determine the difference in each y-value and the one immediately preceding it.

x	y	Δy
−5		
−2		
0		
1		
5		
11		

   f. How can you use the Δy column of the table in *part e* to see that the given points fit a linear function?

   g. Complete the table of values below and, in the last column, determine the difference in each y-value and the one immediately preceding it.

x	y	Δy

−5	**−13**
−2	
1	
4	
7	
10	

h. How can you use the Δy column of the table in *part g* to see that the given points fit a linear function?

i. For any table of values, how can you determine whether the points fit a linear function?

The differences in *y*-values for equally spaced *x*-values are called *first differences*. When first differences are constant, we have constant slope and the data points fit a linear function. This brings up the question of other polynomials. Is there a similar rule for quadratic, cubic, or other polynomial functions? In our study of quadratic functions, we saw that sums of arithmetic series can be expressed using quadratic functions. An arithmetic sequence is formed by starting with some number as first term and then adding the same number, the *common difference*, to form each successive term. When we formed an arithmetic series, we summed up these terms. Thus, for the sum of an arithmetic series, it takes two steps of calculating differences between terms to get back to the common difference of the original arithmetic sequence. Our next goal is to see that there is a similar situation for all quadratic functions. We start on the path to that goal by considering first differences for the general linear function.

4. Every linear function has the form *y* = *ax* + *b*, where *a* and *b* are given real number constants. We'll use a general linear function *y* = *ax* + *b* in each part below.

a. Complete the table of values below, find the differences between successive *y*-values, and explain the result in relation to the formula for the general linear function.

x	$y = ax + b$	First differences
−2		
−1		
0		
1		
2		

b. Complete the table of values below and find the slope between each pair of successive points. What should this value be?

x	$y = ax + b$	Slope between points
−5		
−2		
0		

1		
5		
11		

c. Assume you have a table of values for the general linear function in which the *x*-values are equally spaced 3-units apart, what will be the result of calculating first differences?

d. Assume you have a table of values for the general linear function in which the *x*-values are equally spaced *k*-units apart, what happens with the calculation of first differences?

e. Explain why the points in the following table of values fit a linear function.

x	*y*	First differences
−16	15	
−7	11	
2	7	
11	3	
20	−1	

f. Find the slope of the linear function that fits the points in the table for *part e.*

5. We consider the basic quadratic function $y = ax^2 + bx + c$.
   a. Complete the table of values below. In this table, there is a fourth column for the second differences, which are the differences of the first differences. Calculate the first differences and second differences also.

x	$y = ax^2 + bx + c$	First differences	Second differences
0			
1			
2			
3			
4			

   b. What do you notice about the table in *part a* above?

c. Create another table of values for the general quadratic function $y = ax^2 + bx + c$. This time start with an *x*-value other than 0 and choose *x*-values with a *constant difference* other than 1.

x	$y = ax^2 + bx + c$	First differences	Second differences

d. What happens with the second differences in the table from *part c*?

From our calculations for *Items 4 and 5*, we have seen that, when we start with equally spaced *x*-values (that is, *x*-values that have a constant difference), linear functions have constant *first* differences and quadratic functions have constant *second* differences. We have explored how the constant slope of a line gives the constant first differences for linear functions. We can calculate values of the general quadratic function for inputs of *x, x + k,* and *x + 2k* to prove that the constant second difference for a general quadratic is always $2k^2a$, where *k* is the constant difference in the *x*-values. These results are part of a pattern for polynomial functions. Similar calculations for general polynomials of higher degree show that, when we start with *x*-values that have a constant difference, cubic functions have constant *third* differences, fourth degree polynomials have constant *fourth* differences, and so forth for higher degree polynomials.

.	$y = ax^2 + bx + c$	First differences	Second difference

Least squares regression techniques can be applied to functions and relationships that are not functions. Calculating first differences, second difference, third differences, and so forth to determine whether a polynomial model fits a set of data points applies only to relationships that define functions. When we have a set of data that defines a function, we can apply the method of finite differences to determine whether the data fit a polynomial model.

6. Consider the table of values below.

x	y	
−7	−6	
−5	−5	
−3	3	
−1	22	
1	56	
3	109	
5	185	

a. Explain why it is possible to calculate finite differences for the data points in the table.

b. Determine whether a linear, quadratic, cubic, or other polynomial model fits the data.

	Diameter of viewing circle in centimeters	Area of viewing circle in sq. cms
3	6.	10.5625 π ≈ 33
4	9.0	20.25 π ≈ 64
5	11.5	33.0625 π ≈ 104
6	14.0	49 π ≈ 154
7	16.5	68.0625 π ≈ 214
8	19.0	90.25 π ≈ 284

c. Use the method finite differences to determine if the data points for expressing the diameter of the viewing circles as a function of the distance from the wall fits a polynomial model, and if so, which model.

d. Use the method finite differences to determine if the data points for expressing the area of the viewing circles as a function of the distance from the wall fits a polynomial model, and if so, which model.

e. When the teacher saw this student data, the teacher suspected that the students had just calculated two values and then made up the rest of the data. Why might the teacher have thought this?

7. Return to the data you calculated for the diameter and area of the viewing circle as functions of the distance from the wall.

   a. Use the method of finite differences to determine if your data for the diameter of the viewing circle as a function of the distance from the wall exactly fit a linear model.

Distance from the wall in ft	Diameter of viewing circle in inches	Δy	slope

   b. If a set of data points exactly fit a linear model, what can you say about the least squares regression line?

   c. hat happens with first differences when data points do not exactly fit a linear model but the data can be modeled very well by a regression line?

   d. Use the method of finite differences to determine if your data for the area of the viewing circle as a function of the distance from the wall exactly fit a quadratic model.

   e. If a set of data points exactly fit a quadratic model, what can you say about the least squares regression line?

f. Compare your results with other groups of students in your class. What happens with second differences when data points do not exactly fit a linear model but the data can be modeled very well by a least squares quadratic regression?