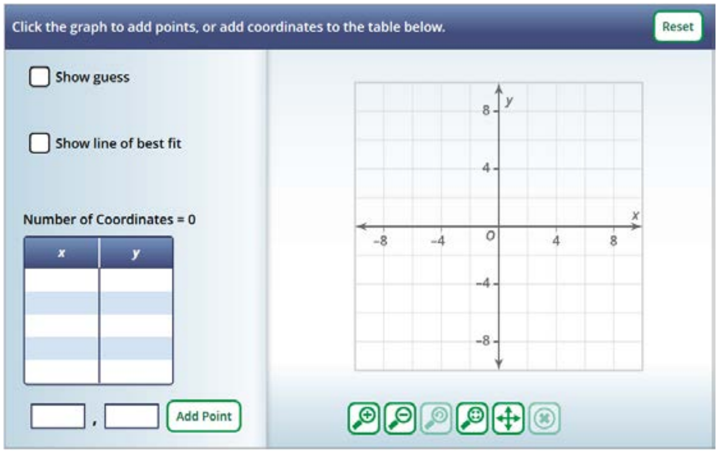
Lesson 14: Modeling Relationships with a Line

Exploratory Activity: Line of Best Fit – Revisited

1. Use the link http://illuminations.nctm.org/Activity.aspx?id=4186 to explore how the line of best fit changes depending on your data set.



   A. Enter any data values for *x* and *y* in the "Add Point" space or simply click on the graph to place a point. Place two points on the grid.

   B. Click on "Show line of best fit" to see where the line is and its equation.

   C. Add more points to your graph to see how the line of best fit changes. You may also drag a point to a new location.

   D. Did anything surprised you when you added points to your graph? Explain.

   E. Try to add points so that your line of best fit changes dramatically. What did you have to do?

   F. Reset the graph and then place points to form a parabola (quadratic function). How well did the line fit this data?

   G. Reset the graph and then place points to form an exponential function. How well did the line fit this data?

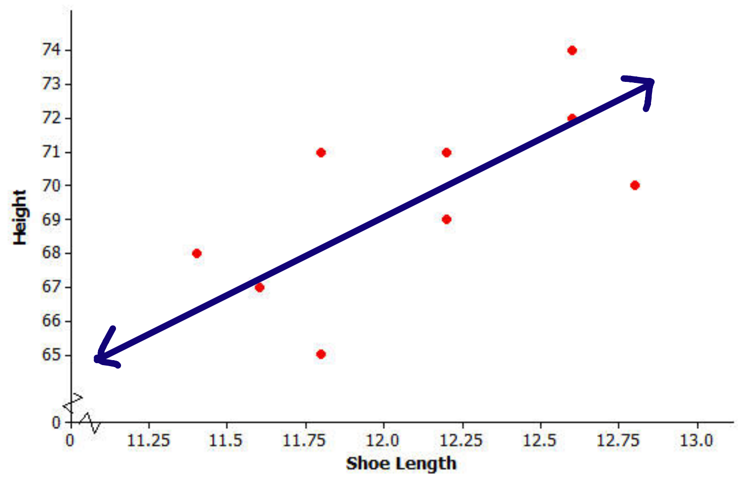
   H. For each line, the program also gave you an "*r*" value. What do you think this *r*-value tells you?

Using a Line to Describe a Relationship

Kendra likes to watch crime scene investigation shows on television. She watched a show where investigators used a shoe print to help identify a suspect in a case. She questioned how possible it is to predict someone's height from his shoe print.

To investigate, she collected data on shoe length (in inches) and height (in inches) from 10 adult men. Her data appear in the table and scatter plot below.

x (Shoe Length)	y (Height)
12.6	74
11.8	65
12.2	71
11.6	67
12.2	69
11.4	68
12.8	70
12.2	69
12.6	72
11.8	71



2. Is there a relationship between shoe length and height? How do you know?

Yes

3. How would you describe the relationship? Do the men with longer shoe lengths tend be taller?

Taller men have larger feet

4. Draw in the line of best fit for this scatterplot. Estimate the height of a man with a shoe length of 12 in.

69 in.

Using Models to Make Predictions

When two variables x and y are linearly related, you can use a line to describe their relationship. You can also use the equation of the line to predict the value of the y-variable based on the value of the x-variable.

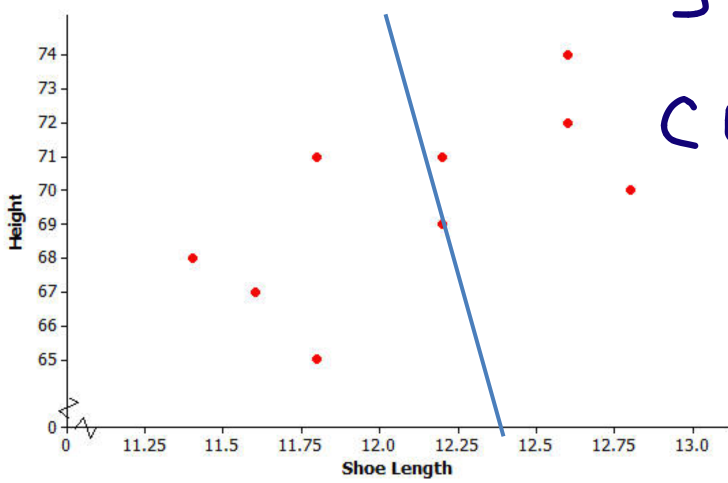
For example, the line $y = 25.3 + 3.66x$ might be used to describe the relationship between shoe length and height, where x represents shoe length and y represents height. To predict the height of a man with a shoe length of 12 in., you would substitute 12 for x in the equation of the line and then calculate the value of y.

5. Use this model to predict a height for a man with a shoe length of 12 in.

$y = 25.3 + 3.66(12) = 25.3 + 43.92 = 69.22$

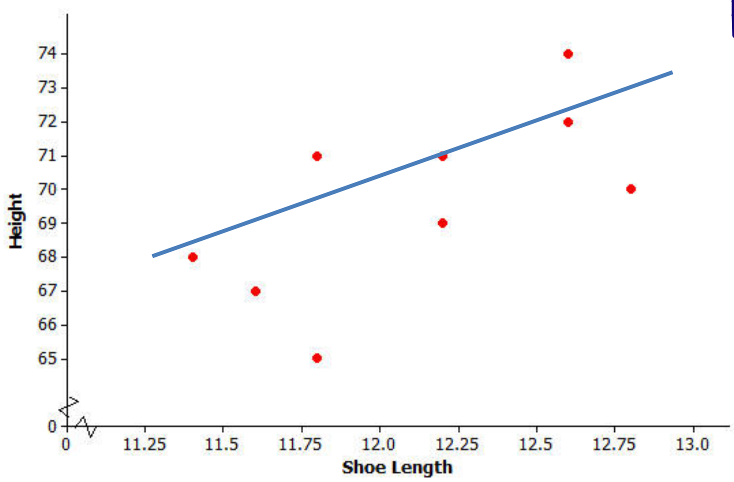
6. Rachel misunderstood how a line of best fit works. Her line is shown below. What might Rachel been considering when she drew this line?

She literally cut the data in half.



7. Josh drew the line of best fit as shown below. Explain to Josh why his line is not appropriate for this data.

His line is too high.



Residuals – How Far Away Are We?

One way to think about how useful a line is for describing a relationship between two variables is to use the line to predict the y-values for the points in the scatterplot. These predicted values could then be compared to the actual y-values.

8. A. For example, the first data point in the table represents a man with a shoe length of 12.6 in. and height of 74 in. Use the line $y = 25.3 + 3.66x$ to predict this man's height.

$y=25.3+3.66(12.6)=25.3+46.116=\boxed{71.416}$

B. Was his predicted height the same as his actual height of 74 in.? Calculate the prediction error by subtracting the predicted value from the actual value. This prediction error is called a ***residual***. For the first data point, the residual is calculated as follows:

Residual $=$ actual y-value $-$ predicted y-value
$= 74 -$ 71.416
$=$ 2.584

9. For the line $y = 25.3 + 3.66x$, calculate the missing values to complete the table.

x (Shoe Length)	y (Height)	Predicted y-value	Residual	Squared
12.6	74	71.42	2.58	6.6564
11.8	65	68.49	−3.49	12.1801
12.2	71	69.95	1.05	1.1025
11.6	67	67.76	−0.76	.5776
12.2	69	69.95	−0.95	.9025
11.4	68	67.02	0.98	.9604
12.8	70	72.15	−2.15	4.6225
12.2	69	69.95	−0.95	.9025
12.6	72	71.42	0.58	.3364
11.8	71	68.49	2.51	6.3001

10. Why is the residual in the table's first row positive and the residual in the second row negative?

The predicted was smaller than the actual, but then the predicted was larger.

11. What is the sum of the residuals? Why did you get a number close to zero for this sum? Does this mean that all of the residuals were close to 0?

−0.6

The positives and negatives cancel.

When you use a line to describe the relationship between two numerical variables, the *best* line is the line that makes the residuals as small as possible overall.

12. If the residuals tend to be small, what does that say about the fit of the line to the data?

Smaller residuals mean the points are closer to the line.

The most common choice for the *best* line is the line that makes the sum of the *squared* residuals as small as possible.

13. Add a column on the right of the table in Exercise 9. Calculate the square of each residual and place the answer in the column.

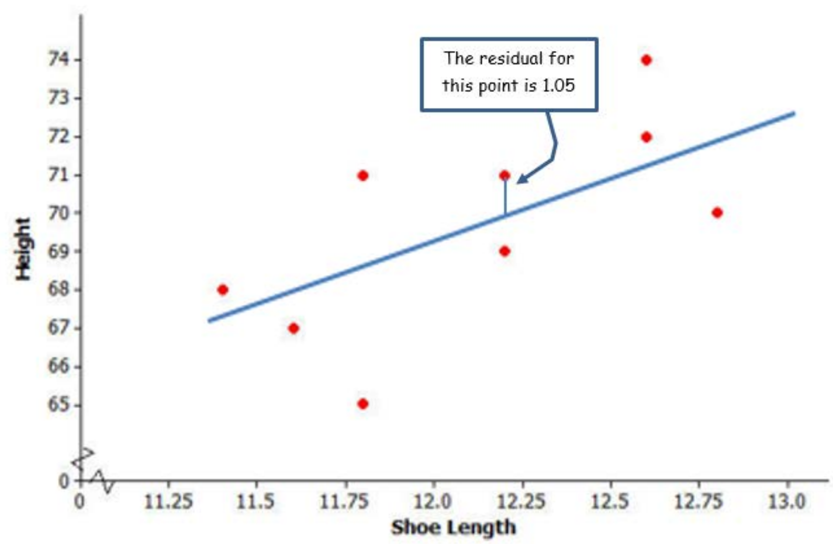
14. Why do we use the sum of the squared residuals instead of just the sum of the residuals (without squaring)? Hint: Think about whether the sum of the residuals for a line can be small even if the prediction errors are large. Can this happen for squared residuals?

So the positives and negatives don't cancel.

15. What is the sum of the squared residuals for the line $y = 25.3 + 3.66x$?

34.54

16. To the right is a graph of the line of best fit and the residuals for this data. Identify the value for each residual.



The Least Squares Line (Best-Fit Line)

The line that has a smaller sum of squared residuals for this data set than any other line is called the *least squares line*. This line can also be called the *best-fit line* or the *line of best fit* (or regression line).

For the shoe-length and height data for the sample of 10 men, the line $y = 25.3 + 3.66x$ is the least squares line. No other line would have a smaller sum of squared residuals for this data set than this line.

There are equations that can be used to calculate the value for the slope and the intercept of the least squares line, but these formulas require a lot of tedious calculations. Fortunately, a graphing calculator can be used to find the equation of the least squares line.

To enter data and obtain the equation of the least squares line using your graphing utility or other statistics program complete the following steps.

Finding the Regression Line (TI-84 Plus)

1. From your home screen, press STAT.
2. From the STAT menu, select the EDIT option. (EDIT, ENTER)
3. Enter the x-values of the data set in L1.
4. Enter the y-values of the data set in L2.
5. Select STAT. Move cursor to the menu item CALC, and then move the cursor to option 4: LinReg($ax + b$) or option 8: LinReg($a + bx$). Press ENTER. (Discuss with students that both options 4 and 8 are representations of a linear equation. Anticipate that most students will be familiar with option 4, or the slope y-intercept form. Option 8 is essentially the same representation using different letters to represent slope and y-intercept. Option 8 is the preferred option in statistical studies.)
6. With option 4 or option 8 on the screen, enter L1, L2, and Y1 as described in the following notes.

LinReg($a + bx$) L1, L2, Y1

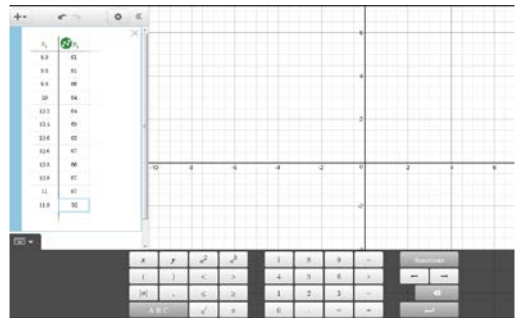
Select ENTER to see results. The least squares regression will be stored in Y1. Work with students in graphing the scatter plot and Y1.

Note: L1 represents the x-values of the regression function, L2 the y-values, and Y1 represents the least squares regression function.

To obtain Y1, go to VARS, move cursor to Y-VARS, and then Functions (ENTER). You are now at the screen highlighting the y-variables. Move cursor to Y1 and hit ENTER.

Y1 is the linear regression line and will be stored in Y1.

Finding the Regression Line (Desmos)



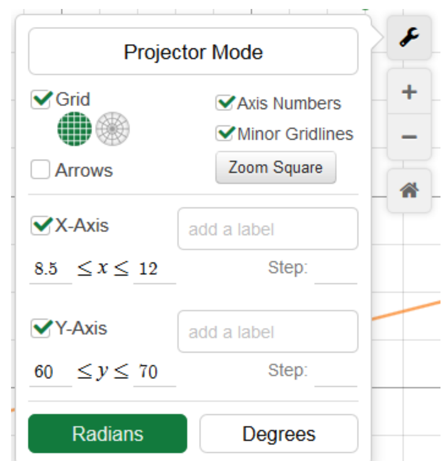
Go to Desmos.com/calculator. You should see a blank grid.

Press the +▾ button and select table.

Enter data in the table.

Press the +▾ button and expression. Then enter y1~mx1+b in the space provided.

Use the wrench tool to adjust the x- and y-axis according to your data.



Press the PLOT button.

The program will automatically graph the line of best fit and give you the slope, y-intercept and list the residuals in the table. The residuals are given as e-values.

16. Enter the shoe-length and height data from Exercise 9 in your graphing utility to find the equation of the least squares line. Round the slope and y-intercept to the nearest hundredth. Did you get $y = 25.3 + 3.66x$?

Yes

17. Assuming that the 10 men in the sample are representative of adult men in general, what height would you predict for a man whose shoe length is 12.5 in.? What height would you predict for a man whose shoe length is 11.9 in.?

$y = 25.3 + 3.66(12.5)$
$y = 25.3 + 45.75$
$\boxed{y = 71.05}$

$y = 25.3 + 3.66(11.9)$
$y = 25.3 + 43.55$
$\boxed{y = 68.85}$

Once you have found the equation of the least squares line, the values of the slope and y-intercept of the line often reveal something interesting about the relationship you are modeling.

The slope of the least squares line is the change in the predicted value of the y-variable associated with an increase of one in the value of the x-variable.

18. Give an interpretation of the slope of the least squares line $y = 25.3 + 3.66x$ for predicting height from shoe size for adult men.

$\frac{3.66}{1}$ $\frac{\text{height (in)}}{\text{shoe length (cm)}}$ For every 1 cm that shoe length increases, height will increase by 3.66 in.

The y-intercept of a line is the predicted value of y when x equals zero. When using a line as a model for the relationship between two numerical variables, it often does not make sense to interpret the y-intercept because a x-value of zero may not make any sense.

19. Explain why it does not make sense to interpret the y-intercept of 25.3 as the predicted height for an adult male whose shoe length is zero.

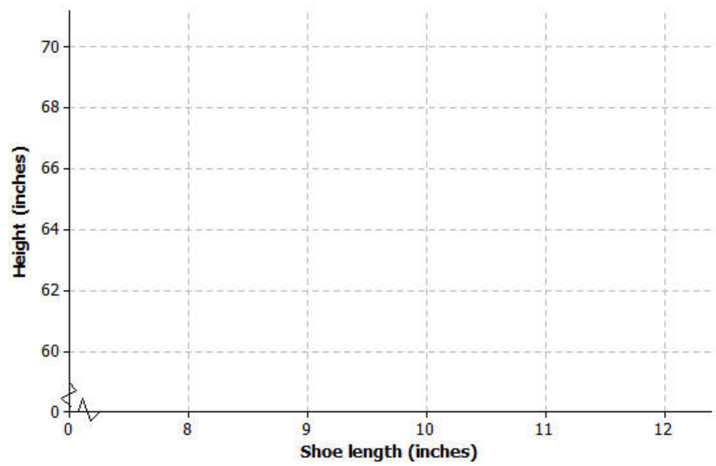
Lesson Summary

- When the relationship between two numerical variables x and y is linear, a straight line can be used to describe the relationship. Such a line can then be used to predict the value of y based on the value of x.
- When a prediction is made, the prediction error is the difference between the actual y-value and the predicted y-value.
- The prediction error is called a *residual*, and the residual is calculated as
  residual $=$ ***actual*** y***-value*** $-$ ***predicted*** y***-value***.
- The *least squares line* is the line that is used to model a linear relationship. The least squares line is the *best* line in that it has a smaller sum of squared residuals than any other line.

Homework Problem Set

Kendra wondered if the relationship between shoe length and height might be different for men and women. To investigate, she also collected data on shoe length (in inches) and height (in inches) for 12 women.

x (Shoe Length of Women)	y (Height of Women)
8.9	61
9.6	61
9.8	66
10.0	64
10.2	64
10.4	65
10.6	65
10.6	67
10.5	66
10.8	67
11.0	67
11.8	70



1. Construct a scatter plot of these data.

2. Is there a relationship between shoe length and height for these 12 women?

3. **GRAPHING UTILITY DEPENDENT:** Find the equation of the least squares line. (Round values to the nearest hundredth.)

4. Suppose that these 12 women are representative of adult women in general. Based on the least squares line, what would you predict for the height of a woman whose shoe length is 10.5 in.? What would you predict for the height of a woman whose shoe length is 11.5 in.?

5. One of the women in the sample had a shoe length of 9.8 in. Based on the regression line, what would you predict for her height?

6. What is the value of the residual associated with the observation for the woman with the shoe length of 9.8 in.?

7. Add the predicted value and the residual you just calculated to the table below. Then, calculate the sum of the squared residuals.

x (Shoe Length of Women)	y (Height of Women)	Predicted Height (in.)	Residual (in.)	Squared Residual
8.9	61	60.72	0.28	
9.6	61	62.92	-1.92	
9.8	66			
10.0	64	64.18	-0.18	
10.2	64	64.81	-0.81	
10.4	65	65.44	-0.44	
10.6	65	66.07	-1.07	
10.6	67	66.07	0.93	
10.5	66	65.76	0.24	
10.8	67	66.7	0.3	
11.0	67	67.33	-0.33	
11.8	70	69.85	0.15	
			Sum of Squared Residuals	

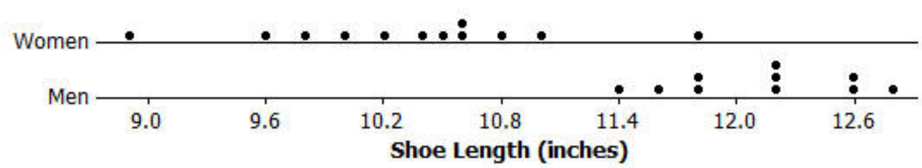
8. Provide an interpretation of the slope of the least squares line.

9. Does it make sense to interpret the y-intercept of the least squares line in this context? Explain why or why not.

10. Would the sum of the squared residuals for the line $y = 25 + 2.8x$ be greater than, about the same as, or less than the sum you computed in Problem 7? Explain how you know this. You should be able to answer this question without calculating the sum of squared residuals for this new line.

11. For the men, the least squares line that describes the relationship between x, which represents shoe length (in inches), and y, which represents height (in inches), was $y = 25.3 + 3.66x$. How does this compare to the equation of the least squares line for women? Would you use $y = 25.3 + 3.66x$ to predict the height of a woman based on her shoe length? Explain why or why not.

12. Below are dot plots of the shoe lengths for women and the shoe lengths for men. Suppose that you found a shoe print and that when you measured the shoe length, you got 10.8 in. Do you think that a man or a woman left this shoe print? Explain your choice.



13. Suppose that you find a shoe print and the shoe length for this print is 12 in. What would you predict for the height of the person who left this print? Explain how you arrived at this answer.