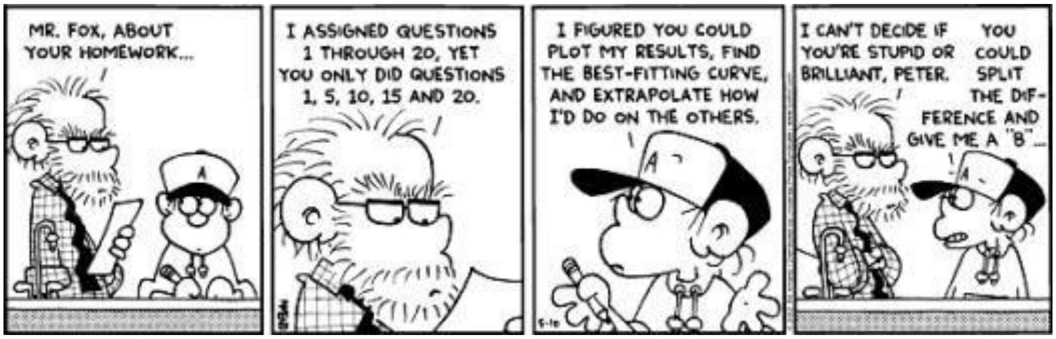
Uncovering Correlation


MR. FOX, ABOUT YOUR HOMEWORK...
I ASSIGNED QUESTIONS 1 THROUGH 20, YET YOU ONLY DID QUESTIONS 1, 5, 10, 15 AND 20.
I FIGURED YOU COULD PLOT MY RESULTS, FIND THE BEST-FITTING CURVE, AND EXTRAPOLATE HOW I'D DO ON THE OTHERS.
I CAN'T DECIDE IF YOU'RE STUPID OR BRILLIANT, PETER.
YOU COULD SPLIT THE DIFFERENCE AND GIVE ME A "B"...

WVU Upward Bound
Summer 2016
June 27 – July 14

Instructor: Jessica Thomas
Email: jmthomas891@gmail.com

Sign up for Remind text messages by texting **@wvuubmath1** to **81010**.

View class resources at **www.symbaloo.com/mix/wvuubmath2016**.

Access class homework by downloading the Socrative Student app or going to **b.socrative.com** and using the room name **MRSTHOMASWV**.

Practice vocabulary words with the Quizlet deck located at **https://quizlet.com/_2dd4xg**.

Name ____

Table of Contents

Symbaloo Legend:

Class resources are located at www.symbaloo.com/mix/wvuubmath2016

- Pink tiles direct to websites used in class activities.
- Purple tiles direct to homework activities, such as Socrative quizzes.
- Red tiles direct to resources for the final project.
- Gray tiles direct to videos viewed in class.
- Blue tiles direct to additional resources to review and practice.

Uncovering Correlation

WVU Upward Bound, Summer 2016
June 27 - July 14

Instructor: Jessica Thomas *Email: jmthomas891@gmail.com*

Course Description

Have you ever wondered if an athlete's height or weight affects their performance? Or if a country's population can have an impact on the number of medals its citizens will win at the Olympics? We are going to discover all that and more throughout this course on correlation. We will create representations of data that will show us how variables affect one another (if at all!) and how strongly they might do so. You will even have the opportunity to research a topic on your own and determine if one variable truly affects another, and how much! When you have finished this course, you will be able to understand and explain the correlation between many real-world variables, including those found at the Olympics!

Unit Objectives

1. Represent data on a scatterplot and describe how the variables are related. (CCSS.MATH.CONTENT.HSS.ID.B.6)
    a. Construct and analyze a scatterplot for correlation.
    b. Construct lines and curves of best fit by eye, by hand, and using technology and use these to interpolate and extrapolate additional values.
    c. Construct and analyze plots of residuals to justify that a line or curve of best fit is accurate.
2. Interpret linear models. (CCSS.MATH.CONTENT.HSS.ID.C)
    a. Calculate the correlation coefficient and coefficient of determination by hand and using technology and use these values to describe a correlation. (CCSS.MATH.CONTENT.HSS.ID.C.8)
    b. Interpret the slope and intercepts of a line of best fit in the context of the situation and data. (CCSS.MATH.CONTENT.HSS.ID.C.7)
    c. Distinguish between correlation and causation. (CCSS.MATH.CONTENT.HSS.ID.C.9)

Class Expectations

1. Mathematics is learned best when it is discovered. You will be told very little in this class, and instead you will uncover a lot of the information for yourself. Have the desire and curiosity to learn new things, and you will succeed!
2. Mathematics is not an individual activity. It requires a lot of teamwork and sharing of ideas. It is expected that you work well with others, share and respect ideas, and help your classmates if they are struggling.
3. Mathematics is cumulative. Everything we will learn is built from something else. Because of this, it is important that you complete all work when it is assigned so that you continue to have success with your learning.

Agenda

Monday	Tuesday	Wednesday	Thursday
June 27: **Welcome!** • Opening Activity • Introductions • Pre-Test • Prerequisite Review	**June 28:** **What's the Correlation?** • Opening Activity • Creating Scatterplots • Studying Correlation • Predict Correlation • Closing Activity	**June 29** **NO CLASS** Upward Bound Fun Day!	**June 30:** **Strength of Correlation** • Opening Activity • Correlation Coefficient Discovery Lab • Calculating the Correlation Coefficient • The Coefficient of Determination • Closing Activity
July 4 **NO CLASS** Fourth of July!	**July 5:** **Lines of Best Fit "by Eye"** • Opening Activity • Defining a Line of Best Fit • Writing Equations • Interpreting the Line of Best Fit • Making Predictions • Closing Activity	**July 6:** **Residuals and Least Squares Regression** • Opening Activity • Defining Residuals • Residuals Discovery Lab • Least Squares Simulation • Closing Activity	**July 7:** **Linear and Nonlinear Regression** • Opening Activity • Finding the Least Squares Regression Line • Linear Regression Using Excel • Linear Regression Using Desmos • Nonlinear Regression Discovery Lab • Closing Activity
July 11: **The 'Ations** • Opening Activity • Interpolation vs. Extrapolation • Correlation and Causation • Reasons for Correlation • Closing Activity	**July 12:** **Partner Projects** • Opening Activity • Data Gathering • Constructing Scatterplots • Calculating Correlation • Line of Best Fit • Closing Activity	**July 13:** **Partner Projects** • Line of Best Fit • Residuals • Nonlinear Regression Test • Test for True Correlation • Creating PowerPoints • Presentations	**July 14:** **Partner Projects** • Presentations • Post-Test

Ingenuity + Courage + Work = MIRACLES

– Bob Richards, Olympic Gold Medalist

Name ____

What's the Correlation?

We now have three different scatterplots. Sketch them out here.

What do you notice about each of these scatterplots?

How would you explain what is happening in each of these scatterplots?

Definitions:

- independent variable:

- dependent variable:

- positive correlation:

- negative correlation:

Enrichment:

Can you come up with three unique situations about the Olympics that would have data that would produce scatterplots with each of the three correlations?

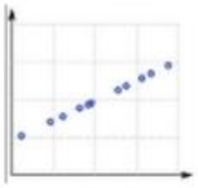
Discovering the Correlation Coefficient

Phet Simulation

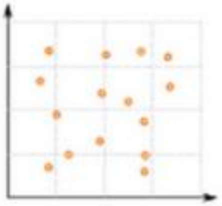
To begin, go to the simulation using this link: http://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html, or click the pink tile on the Symbaloo.

On the top drop-down menu, select "Custom." For each of the following scatterplots, try and duplicate the plots on your screen. You can grab data points from the bowl in the bottom left side of your screen. Matching the actual values is not of the greatest importance; you are simply trying to duplicate the arrangement of the points. You will want to pay close attention to when the points are clustered close together or are farther apart.

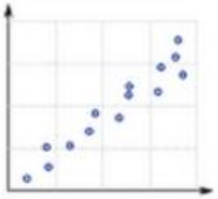
After you have created each plot, click the green plus sign on the left side to reveal the correlation coefficient. Record this value (r = ?).



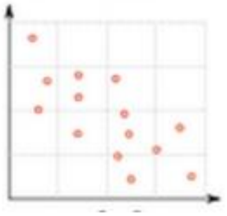
r =



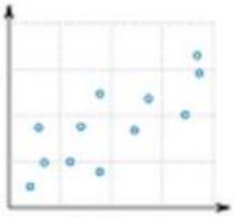
r =



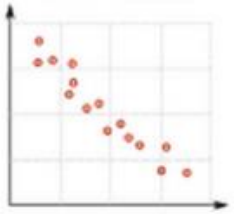
r =



r =



r =



r =

Can you construct a scatterplot with a correlation coefficient greater than 1?

Can you construct a scatterplot with a correlation coefficient of 0?

Can you construct a scatterplot with a correlation coefficient less than −1?

Name ____

Analysis

Based on your work in the simulation, answer these analysis questions.

1. Next to each of the graphs, write whether the scatterplot shows a positive, negative, or no correlation.

2. Write a sentence or two summarizing any connections you can note between the r-value and the type of correlation you recorded for question #1.

3. What range can the r-value have?

4. What does an r-value close to positive 1 seem to mean?

5. What does an r-value close negative 1 seem to mean?

6. What does an r-value close to 0 seem to mean?

Enrichment

7. Construct a scatterplot with 10-20 data points. Estimate the correlation coefficient. How close were you?

8. Create a new plot in the phet simuation using only two points. What is your r-value? Will your r-value always be this number if you have two points? Explain.

Calculating the Correlation Coefficient

Calculating by Hand – Best with Small Data Sets!

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n(\sum x^2) - (\sum x)^2)\,(n(\sum y^2) - (\sum y)^2)}}$$

This formula looks really scary, but it is just multiplication, addition, and subtraction!

First, estimate your correlation coefficient based on your scatterplot. This makes it simple to double check your answer! To make using the formula easier, fill out a table like this one.

n = number of data points =					
	x	y	xy	x^2	y^2
Sum ($\sum$)	$\sum x =$	$\sum y =$	$\sum xy =$	$\sum x^2 =$	$\sum y^2 =$
Square (^{2})	$\left(\sum x\right)^2 =$	$\left(\sum y\right)^2 =$			

Then, simply input the values and solve!

$$r = \frac{(\qquad)(\qquad) - (\qquad)(\qquad)}{\sqrt{((\qquad)(\qquad) - (\qquad))((\qquad)(\qquad) - (\qquad))}}$$

Calculating with Technology

Fortunately, you can calculate the correlation coefficient with technology much easier.

1. Open an Excel spreadsheet.
2. Enter your x values in column A, and enter your y values in column B.
3. Click in an open cell. In the formula bar, type **=CORREL(A:A,B:B)** and hit Enter. The actual correlation coefficient should appear in the cell. (Data must be in columns A and B for this formula to work.)

Let's try this with the data above. Were our calculations correct?

Definitions:

- correlation coefficient:
  - characteristics of the correlation coefficient:
- variation:

Enrichment

The coefficient of determination assesses how well a model explains a situation. The coefficient of determination tells you what percent of the variation in the dependent variable (y) is explained by the variation in the independent variable (x). For example, if the coefficient of determination is 0.5, then 50% of the variation is due to the variation in the independent variable. The coefficient of determination is calculated by squaring the correlation coefficient and is denoted r^2.

1. If the correlation coefficient is 0.87, what is the coefficient of determination? Explain what this value means in words.
2. What range can the coefficient of determination have?
3. What would be some ideal values for a coefficient of determination?

Creating a Line of Best Fit

Use the space below to record the rules that we write about creating and defining lines of best fit.

Rule	Example	Counterexample?

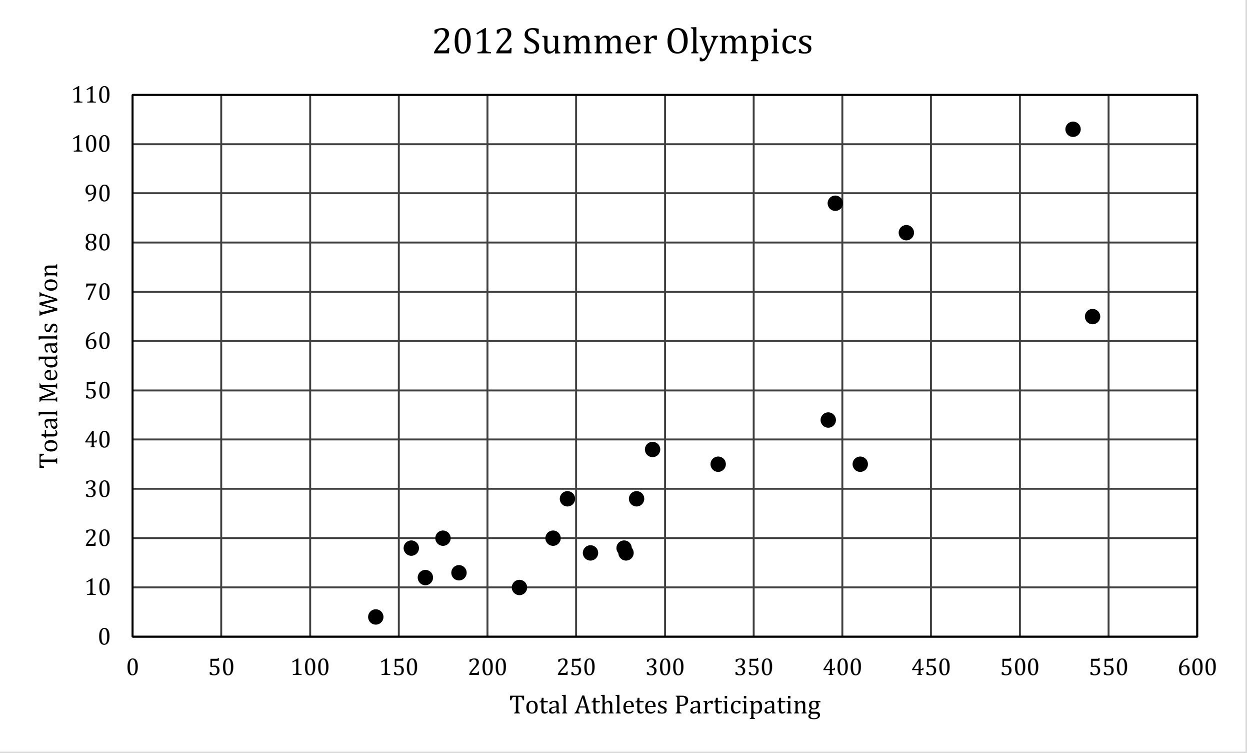
Name ____

Writing and Explaining an Equation for Line of Best Fit

The following data and scatterplot is from the 2012 Summer Olympics in London.

Country	Total Athletes Participating	Total Medals Won
Great Britain	541	65
United States	530	103
Russia	436	82
Australia	410	35
China	396	88
Germany	392	88
France	330	35
Japan	293	38
Italy	284	28
Spain	278	17

Country	Total Athletes Participating	Total Medals Won
Canada	277	18
Brazil	258	17
South Korea	245	28
Ukraine	237	20
Poland	218	10
New Zealand	184	13
Netherlands	175	20
Belarus	165	12
Hungary	157	18
Argentina	137	4



Make an observation about the data. What predictions or assumptions can you make regarding total athletes and total medals?

Name ____________________

Thinking about the Data

1. Sketch a line of best fit for the data on the scatterplot.
2. When a country sends 50 more athletes than another country, how many more medals do they seem to win? Write this as a ratio of more medals:more athletes.
3. When a country sends 100 more athletes than another country, how many more medals do they seem to win? Write this as a ratio of more medals:more athletes.
4. When a country sends 200 more athletes than another country, how many more medals do they seem to win? Write this as a ratio of more medals:more athletes.
5. If a country wins 0 medals, predict how many athletes they sent to the Olympics.
6. If a country sends 0 athletes, how many medals does your line say they should win? Is this a valid amount of medals? Why or why not?

Writing an Equation for the Line of Best Fit

7. What is the slope of your line? You can estimate. (It might be helpful to refer to questions 2-4).
8. What is the approximate x-intercept of your line? (It might be helpful to refer to question 5.)
9. What is the approximate y-intercept of your line? (It might be helpful to refer to question 6).

10. What is the equation for your line?

11. Explain the slope of your line in the context of the situation. What happens to the dependent variable when the independent variable increases?

12. Explain the y-intercept of your line in the context of the situation. What is the dependent variable equal to when the independent variable is 0?

Making Predictions

13. You are the president of another country and want to enter the 2016 Olympics. Based on the information from the 2012 Olympics and the line of best fit you've created, how many athletes should you send if you want to win 45 medals?

14. A rival country has 500 athletes to send to the 2016 Olympics. Who will win the most medals: you or the new rival country? By how many medals?

Enrichment

Why might some of the points be more spread out than others (see the points on the right side of the scatterplot)?

Line of Best Fit Guessing Game

Go to: http://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html or click on the pink tile in the Symbaloo. Choose a topic from the drop-down bar and see how close you can get to the real line of best fit! For predictions, choose any value of x that you like and/or are given.

Scatterplot Title: ______________________________

My estimated best fit line: ______________________________

Actual best fit line: ______________________________

Explain the slope: When ____________________ increases/decreases by __________

the ____________________ increases/decreases by __________

Predict: When x = ____________, y = ____________.

Scatterplot Title: ______________________________

My estimated best fit line: ______________________________

Actual best fit line: ______________________________

Explain the slope: When ____________________ increases/decreases by __________

the ____________________ increases/decreases by __________

Predict: When x = ____________, y = ____________.

Scatterplot Title: ______________________________

My estimated best fit line: ______________________________

Actual best fit line: ______________________________

Explain the slope: When ____________________ increases/decreases by __________

the ____________________ increases/decreases by __________

Predict: When x = ____________, y = ____________.

Name ____

Scatterplot Title: ____

My estimated best fit line: ____

Actual best fit line: ____

Explain the slope: When _______________________ increases/decreases by ________

the _________________________ increases/decreases by ________

Predict: When x = ____________, y = ____________.

Scatterplot Title: ____

My estimated best fit line: ____

Actual best fit line: ____

Explain the slope: When _______________________ increases/decreases by ________

the _________________________ increases/decreases by ________

Predict: When x = ____________, y = ____________.

Scatterplot Title: ____

My estimated best fit line: ____

Actual best fit line: ____

Explain the slope: When _______________________ increases/decreases by ________

the _________________________ increases/decreases by ________

Predict: When x = ____________, y = ____________.

Exploring Residuals

Amber, Brandon, and Cody have each created a line of best fit for the data below. Now it's time to decide which is the best one. Their teacher has explained the concept of residuals to them, and they want you to test each of their lines to see who constructed the most accurate line of best fit.

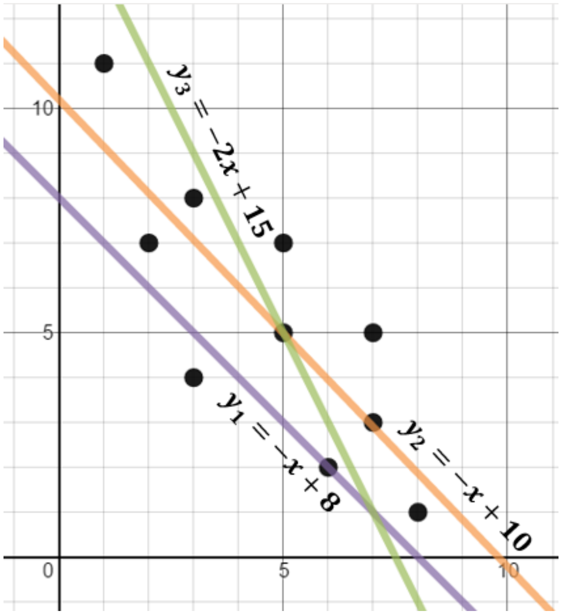
A residual is ____

____

You can calculate a residual by ____

____

x	y
1	11
2	7
3	8
3	4
5	7
5	5
6	2
7	3
7	5
8	1



Amber's equation:
$y_1 = -x + 8$

Brandon's equation:
$y_2 = -x + 10$

Cody's equation:
$y_3 = -2x + 15$

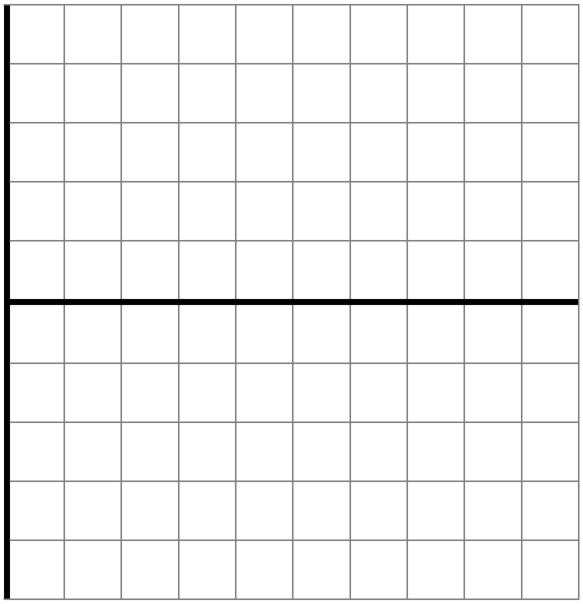
Which line of best fit do you think is the most accurate? Why?

Name ____

First, test Amber’s equation by completing the chart below.

x	Observed y	Predicted y_1 $y_1 = -x + 8$	Residual: $y_1 - y$
1	11		
2	7		
3	8		
3	4	$-(3) + 8 = 5$	$5 - 4 = 1$
5	7		
5	5		
6	2		
7	3		
7	5		
8	1		

Then, plot the residual values as points on the graph below. The residual point for the data point (3,4) will be (4,1) because it is the 4th data point and has a residual value of 1.

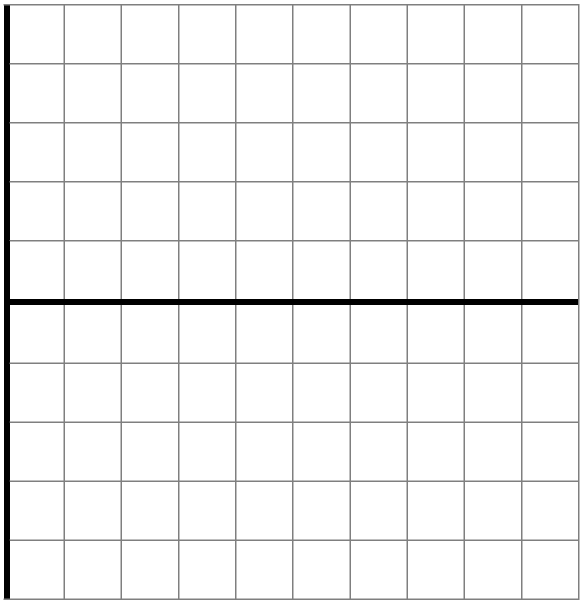


Lastly, calculate the average of the residuals for Amber’s line of best fit.

Now test Brandon’s equation by completing the chart below.

x	Observed y	Predicted y_1 $y_2 = -x + 10$	Residual: $y_2 - y$
1	11		
2	7		
3	8		
3	4	$-(3) + 10 = 7$	$7 - 4 = 3$
5	7		
5	5		
6	2		
7	3		
7	5		
8	1		

Then, plot the residual values as points on the graph below. The residual point for the data point (3,4) will be (4,3) because it is the 4th data point and has a residual value of 3.

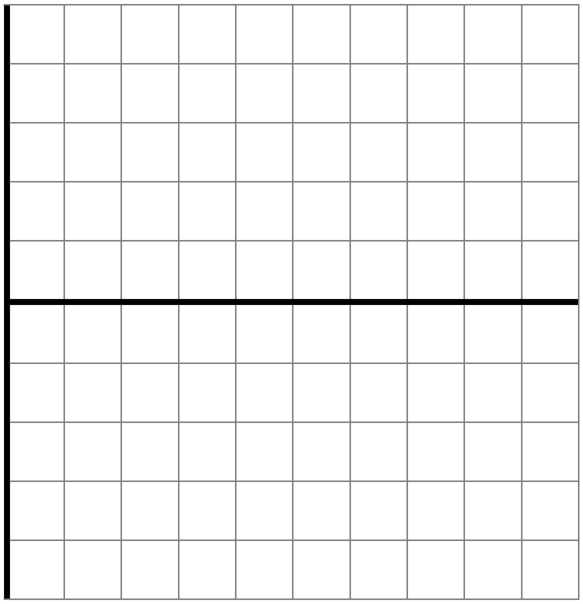


Lastly, calculate the average of the residuals for Brandon’s line of best fit.

Last, test Cody’s equation by completing the chart below.

x	Observed y	Predicted y_1 $y_3 = -2x + 15$	Residual: $y_3 - y$
1	11		
2	7		
3	8		
3	4	$-2(3) + 15 = 9$	$9 - 4 = 5$
5	7		
5	5		
6	2		
7	3		
7	5		
8	1		

Then, plot the residual values as points on the graph below. The residual point for the data point (3,4) will be (4,5) because it is the 4th data point with a residual value of 5.



Lastly, calculate the average of the residuals for Cody’s line of best fit.

Analysis

Investigate all three residuals plots and residuals averages side-by-side.

It is fairly clear that Amber's line of best fit is a little too low for the data set.

1. What does this "too low" line do to her residual plot compared to the residual plots for the other lines of best fit?

2. What does this "too low" line do to the average of her residuals compared to the other averages?

It's hard to tell if Brandon or Cody's lines of best fit are the most accurate for the data set. However, you know that the line should come as close to possible to all data points.

3. For the line to come as close as possible to all data points, what should the average of the residuals be closest to?

4. Based on your answer for question #3, who has the most accurate line of best fit?

5. Brandon's line of best fit produces a scatterplot with what type of correlation?

6. Cody's line of best fit produces a scatterplot with what type of correlation?

7. What type of correlation should a residual plot have if the line of best fit is as accurate as possible?

Residuals and Least Squares Regression Notes

Definitions:

- residual:

    - how to calculate a residual:

    - characteristics of a residuals plot:

    - average of residuals:

- least squares regression:

Why are residuals squared before adding them in least squares regression?

Why does the sum of squares have to be as small as possible in least squares regression?

Name ____

Calculating the Least Squares Regression Line

By Hand

The calculation for finding the equation of the least squares regression line is very similar to calculating the correlation coefficient.

$$y = \frac{n(\sum xy) - \sum x \sum y}{n(\sum x^2) - (\sum x)^2} x + \frac{\sum x^2 \sum xy - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

Complete a table to find all of the values you will need.

n = number of data points =					
	x	y	xy	x^2	y^2
Sum (∑)	$\sum x =$	$\sum y =$	$\sum xy =$	$\sum x^2 =$	$\sum y^2 =$
Square (^{2})	$\left(\sum x\right)^2 =$	$\left(\sum y\right)^2 =$			

Then input the values and solve!

$$y = \frac{(\quad)(\quad) - (\quad)(\quad)}{(\quad)(\quad) - (\quad)} x + \frac{(\quad)(\quad) - (\quad)(\quad)}{(\quad)(\quad) - (\quad)}$$

With Technology

Using Excel

1. Open Excel.
2. In cell A1, type the name of the first variable. In cell B1, type the name of the second variable.
3. List the data points in the columns. The order does not matter.
4. Highlight all of your data.
5. Click on the Insert tab at the top.
6. Select the Scatterplot option from the middle of the ribbon, then select the first scatterplot option. Your scatterplot should be created automatically for you.
7. Click on your scatterplot and then click on the Design tab under "Chart Tools" at the top.
8. Click on "Add Chart Element" on the left-hand side. Select "Trendline" and then "Linear." A trendline, or line of best fit, should appear on your scatterplot.
9. Right click on the trendline and select "Format Trendline."
10. A toolbar should appear on the right side. Click on the third icon, the bar graph. Scroll to the bottom and check "Display equation on chart." The equation of the line should appear on the scatterplot.

Using Desmos

1. Go to www.desmos.com/calculator, click on the pink tile on the Symbaloo, or download and open the Desmos app on your device.
2. Click on the plus sign in the top left and choose "table." The table should load into Box 1.
3. Enter your values in this table and they should appear automatically on the graph.
4. Click into Box 2.
5. To do a linear regression, type "y1~mx1+b". (If the top of your table from Step 3 reads something other than x_1 and y_1, you will have to change the values in the equation accordingly.)
6. Box 2 should now contain data about your line of best fit, such as the correlation coefficient and parameters (slope and y-intercept).
7. To check the residual plot, click the "plot" button in box 2. The residuals will appear near the x-axis, and the residual values will appear in another column in the table in Box 1.

Desmos Tips: If you want to change the color of the points or the line to make it easier to see, simple click and hold the colored circles and more color options will appear. To hide a line or set of points, simply click off its colored circle. To make points and lines easier to see, click the wrench in the top right and turn on Projector Mode.

Nonlinear Regression: When Linear Regression Doesn't Cut It

Amber, Brandon, and Cody have a new data set and are trying to create the most accurate regression possible, and they need your help determining who is right – again.

The data below depict the number of women athletes at each of the 28 Summer Olympics since 1896 (minus 1940 and 1944, when the Olympics were canceled due to WWII).

Olympic Number	# of Women Athletes
1	0
2	23
3	6
4	6
5	44
6	53
7	78
8	156
9	312
10	202
11	361
12	446
13	521
14	371
15	613
16	680
17	783
18	1060
19	1260
20	1123
21	1567
22	2202
23	2721
24	3520
25	4068
26	4303
27	4611
28	4655

Enter this information into a table in Desmos (www.desmos.com/calculator). See your notes on how to do so if you have forgotten.

Does this data look linear? ________________

Create a linear regression for this data (see your previous notes).

What is the equation for the line that models this data? ______________________________________

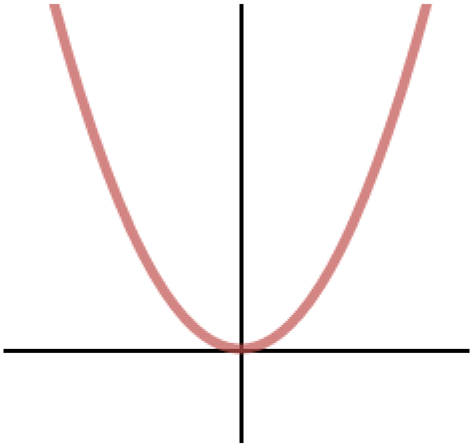
$r =$ _________________________ $r^2 =$ ________________________________

Turn on the residuals for this line. Sketch the plot below. Is this a good residuals plot? Why or why not?

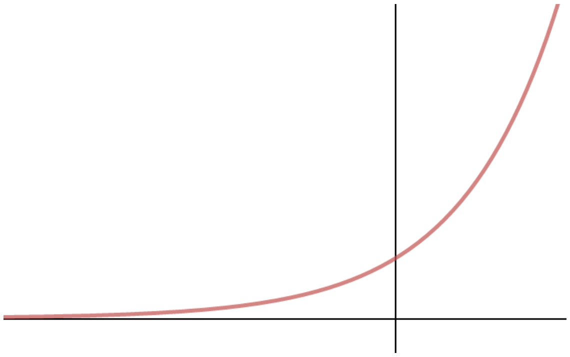
Do you think that this data can be accurately modeled with a linear regression?

Amber, Brandon, and Cody decided that a line of best fit wasn't appropriate for this data set. They looked over the different parent functions they know, and each picked a different one that they think will produce the most accurate curve of best fit.

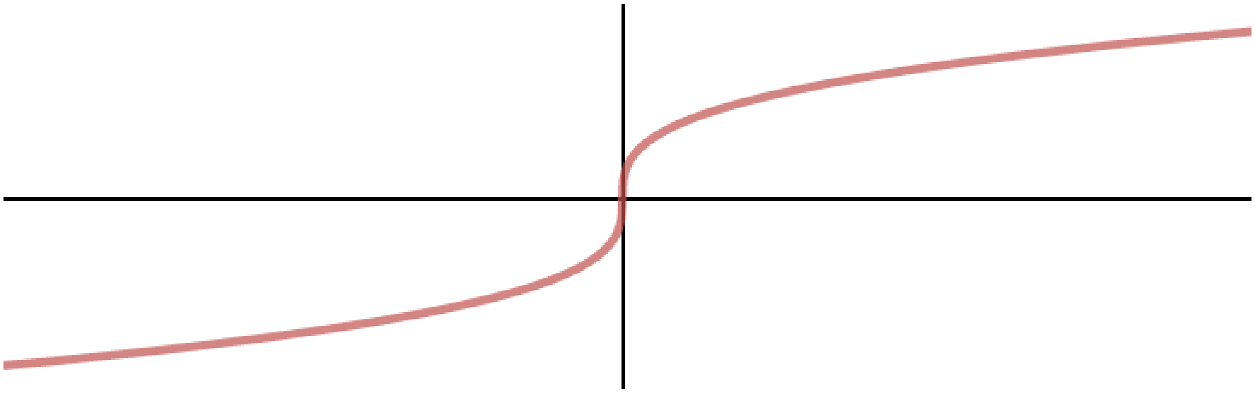
Amber thinks that the data looks like it models the right side of quadratic function ($y = x^2$)



Brandon thinks that the data looks like it models an exponential function ($y = b^x$)



Cody thinks that the data looks like it models a cube root function ($y = \sqrt[3]{x+b} + c$)



Who do you think is right? Why?

Name ____

Test Amber's theory using Desmos. Leaving the same data set in the table, graph a new regression line by typing "y1~ax1^2" to produce a quadratic regression line.

What is the equation for the quadratic regression line? ____________________________________

$r^2 =$ _______________________

Turn on the residuals for this line. Sketch the plot below. Is this a good residuals plot? Why or why not?

Do you think that this data can be accurately modeled with a quadratic regression?

Test Brandon's theory using Desmos. Leaving the same data set in the table, graph a new regression line by typing "y1~b^(x1+a)" to produce an exponential regression line.

What is the equation for the exponential regression line? __________________________________

$r^2 =$ _______________________

Turn on the residuals for this line. Sketch the plot below. Is this a good residuals plot? Why or why not?

Do you think that this data can be accurately modeled with an exponential regression?

Test Cody's theory using Desmos. Leaving the same data set in the table, graph a new regression line by typing "y1 = a nthroot 3 (x1+b) + c to produce a cube root regression line. (This one can be difficult to type in. Ask for help if you need it.)

What is the equation for the cube root regression line? ________________________________

$r^2 =$ ____________________

Turn on the residuals for this line. Sketch the plot below. Is this a good residuals plot? Why or why not?

Do you think that this data can be accurately modeled with a cube root regression?

Which regression most accurately models the data on these female Olympic athletes? How did you make your decision?

Use the model to predict how many women athletes will be at the 2016 Rio Olympics (Olympics #29).

Name ____

Summarizing the Investigation

A line of best fit, when data is strongly correlated, should have a correlation coefficient *r* as close as possible to _________ or ___________and a coefficient of determination r^2 as close as possible to __________

A curve of best fit cannot have a correlation coefficient, but its coefficient of determination r^2 should still be as close as possible to __________.

A line of best fit should produce a residuals plot with _________________________ correlation, and the average of residuals should be as close as possible to _________.

A curve of best fit should also produce a residuals plot with _________________________ correlation, and the average of residuals should also be as close as possible to ______________

Why is it important to construct a curve of best fit if a line of best fit does not provide a good coefficient of determination or residuals plot?

Enrichment

What other types of functions might model data? How would you write an equation to test these in Desmos?

Name ____

[Page intentionally left blank]

The 'Ations –Definitions and Notes

Definitions:

- interpolation:
- extrapolation:
- confounding variable:
- causal mechanism:

Explain the phrase "correlation does not imply causation."

When two variable are correlated, they can be written ____

An example of this is:

Four reasons for looking for correlation:

1.
2.
3.
4.

Correlation and Causation Video Notes

Record the information from the studies presented in the HowStuffWorks video.

1. Independent variable ______________________________

   Dependent variable ______________________________

   Confounding variable ______________________________

2. Independent variable ______________________________

   Dependent variable ______________________________

   Confounding variable ______________________________

3. Independent variable ______________________________

   Dependent variable ______________________________

   Confounding variable ______________________________

4. Independent variable ______________________________

   Dependent variable ______________________________

   Confounding variable ______________________________

5. Independent variable ______________________________

   Dependent variable ______________________________

   Confounding variable ______________________________

Are there any other confounding variables you can think of for these situations?

Spurious "Correlations"?

Go to: http://tylervigen.com/old-version.html, or click on the pink tile on the Symbaloo.

Find three (or more!) different scatterplots/graphs depicting different correlations. There are 25,000 graphs on the site – dig deep to find ones that interest you!

1. Independent variable: ____

   Dependent variable: ____

   Is this just a random correlation or a true correlation with causation? Explain your thinking. What might be the confounding variable? What might be the causal mechanism? ____

   ____

   ____

   ____

2. Independent variable: ____

   Dependent variable: ____

   Is this just a random correlation or a true correlation with causation? Explain your thinking. What might be the confounding variable? What might be the causal mechanism? ____

   ____

   ____

   ____

3. Independent variable: ____

   Dependent variable: ____

   Is this just a random correlation or a true correlation with causation? Explain your thinking. What might be the confounding variable? What might be the causal mechanism? ____

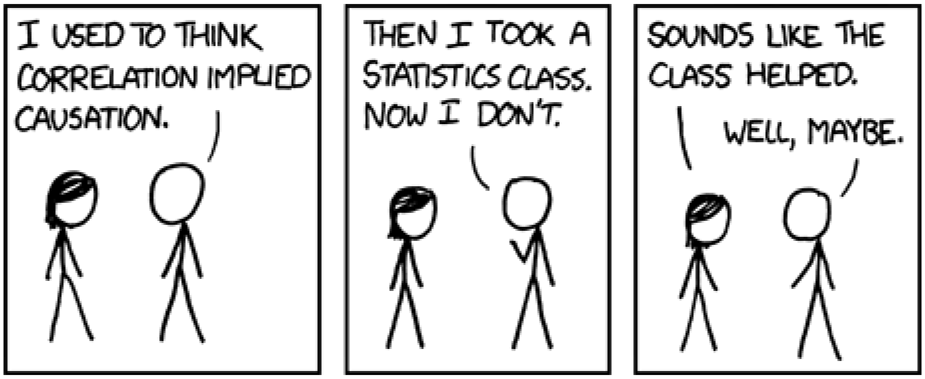
   ____

   ____

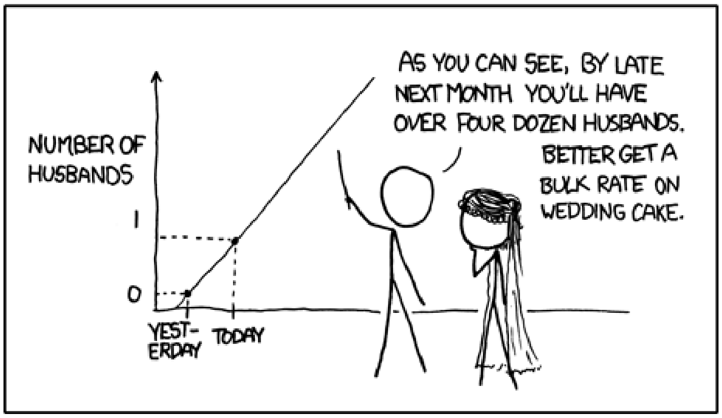
   ____

What's the Joke?

Explain the joke in each of the comics below.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

MY HOBBY: EXTRAPOLATING


NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

Uncovering Correlations in the Olympics

Task:

For this project, you and a partner will analyze the relationship between two variables of your choosing related to the Summer Olympics. The first steps will be gathering the data, creating a scatterplot, and finding a line or curve of best fit for the data. Then, use this regression function to make predictions and explain the situation. Determine how well your variables are correlated by calculating the correlation coefficient, coefficient of determination, and residuals. Consider your data and results in terms of correlation and causation to determine if one causes the other. After you have thoroughly investigated the variables, prepare a presentation to share with the rest of the class!

Sample Variables/Topics:

- Country wealth or population
- Number of medals won (at one Summer Games or all Summer Games)
- Number of athletes sent (men, women, or both)
- Ages, heights, or weights of athletes
- Winning times or scores from different events
- Number of sports events participated in

Helpful Resources (located on red tiles on the Symbaloo):

- http://www.sports-reference.com/olympics/
- https://www.olympic.org/olympic-results
- Wikipedia pages can provide information about athletes and countries

Possible Presentation and Technology Tools:

- Excel
- Desmos
- PowerPoint
- Google Slides
- Prezi

Name ____

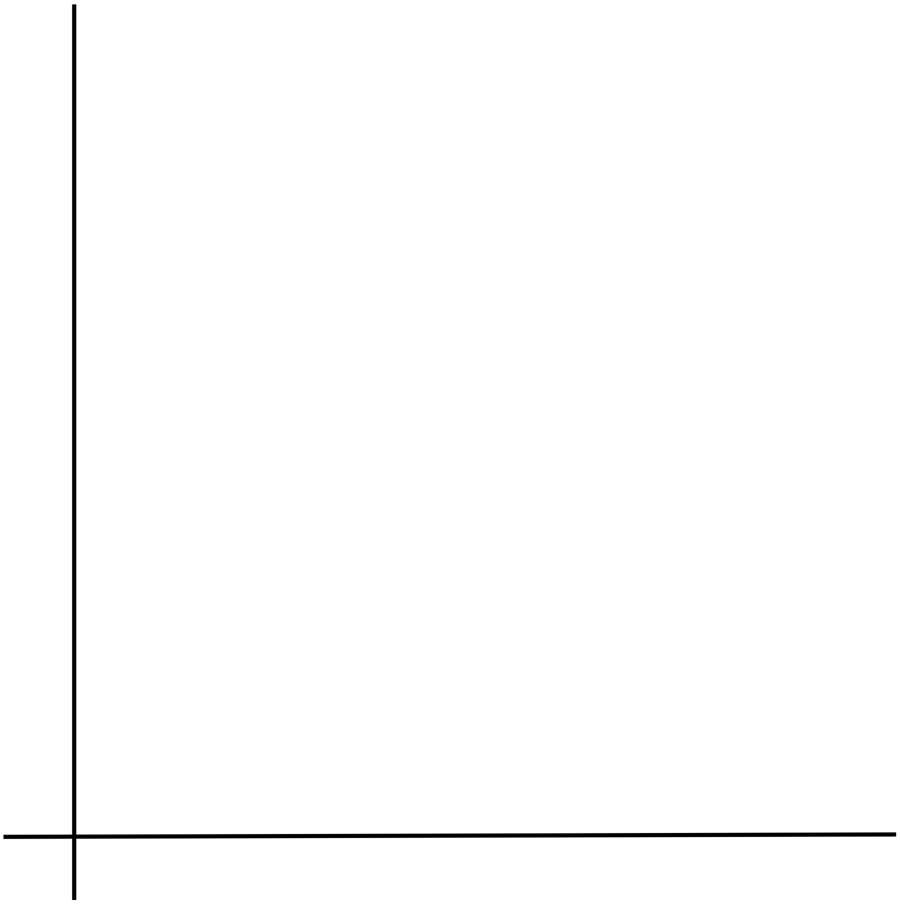
Scoring Rubric:

Task	*Possible Score*	*Total Score for Stage*	*Earned Score*
Stage 1: Constructing the Scatterplot		**22**	
Identify independent and dependent variable	2		
Collect data (must have at least 20 data points unless teacher has approved otherwise)	10		
Construct scatterplot of data (using Excel or Desmos)	10		
Stage 2: Finding Correlation		**18**	
Record type of correlation (positive, negative, no)	5		
Calculate correlation coefficient and coefficient of determination of data	5		
Explain correlation coefficient and coefficient of determination	8		
Stage 3: Finding the Best Fit Line or Curve		**25**	
Calculate line or curve of best fit	10		
Explain line or curve of best fit in the context of the situation	10		
Interpolate two additional values and extrapolate two additional values using the line or curve of best fit	5		
Stage 4: Checking Residuals		**20**	
Calculate residuals	10		
Plot residuals	8		
Average residuals	2		
Stage 5: Explaining Correlation		**10**	
Explain the correlation between the two variables using correct vocabulary	5		
Decide if a causation exists and explain what it might be	5		
Stage 6: Presenting		**5**	
Construct a presentation containing all information that is relatively free of errors and easy to understand	2		
Present research clearly and answer any questions	2		
Cite all sources	1		
Total		**100**	

Stage 1: Constructing the Scatterplot

- Select two variables to investigate. Have your variables approved by the teacher. If you have trouble selecting variables, you may choose from the provided list.
- You must be able to collect at least 20 data points.
- Record your data using the table below. Create a scatterplot using technology and draw a rough sketch of the scatterplot on the axes. Remember to use all unit and graph labels.

Independent Variable x:	Dependent Variable y:



Name ____

Stage 2: Finding Correlation

- What correlation is evident in your data (positive, negative, no)?
- Calculate the correlation coefficient.
- What does this correlation coefficient mean?
- Calculate the coefficient of determination.
- What does this coefficient of determination mean?

Stage 3: Finding the Best Fit Line or Curve

- Find the line of best fit or curve of best fit that most accurately describes the data. Try many different functions to find the best one. Sketch it on your scatterplot on the previous page. What is the equation for this regression?
- Explain this line or curve in the context of the variable you are investigating.
- Predict four additional values using this line or curve. Extrapolate two values and interpolate two values.

** Remember, if you chose a curve of best fit, you may have to go back and change your coefficient of determination and eliminate your correlation coefficient!

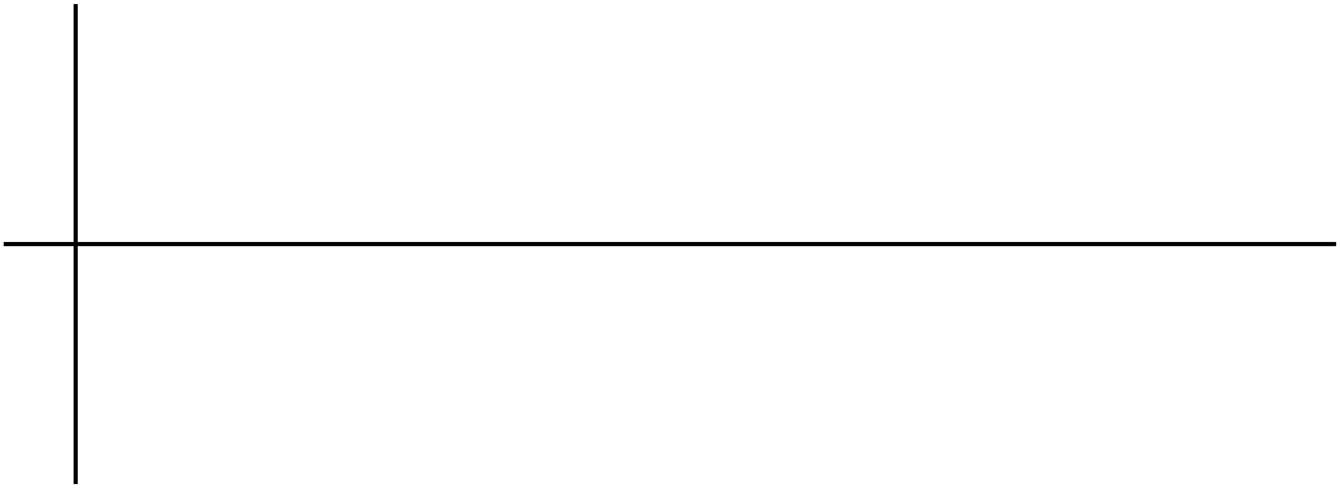
Stage 4: Calculating Residuals

- Calculate the residuals of your data. You might need more rows.

x	Observed y	Predicted y_1 $y_1 =$	Residual: $y_1 - y$

- What is the average of your residuals?

- Plot your residuals on the axes below.



Stage 5: Explaining the Correlation

- Based on your work in Stages 1-4, explain the correlation evident between the variables you investigated. What evidence supports that this correlation exists? Be very explicit and use correct vocabulary!

- Do you think this is an example of correlation with causation? If so, what might the causal mechanism be? If not, what might the confounding variable be?

Stage 6: Presenting

- Put all of your work on this investigation into a presentation tool such as PowerPoint, Google Slides, or Prezi. Prepare to present to your peers and answer any questions!