Lesson 7: Modeling a Context from Data

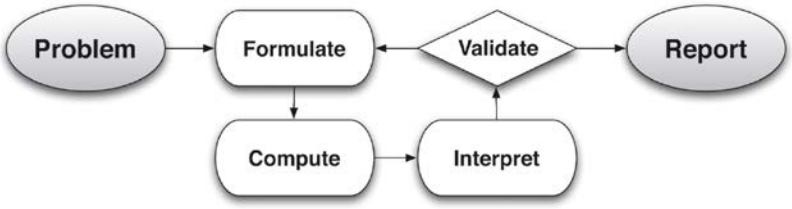
Student Outcomes

- Students use linear, quadratic, and exponential functions to model data from tables and choose the regression most appropriate to a given context. They use the correlation coefficient to determine the accuracy of a regression model and then interpret the function in context. They then make predictions based on their model and use an appropriate level of precision for reporting results and solutions.

Lesson Notes

Lesson 7 focuses on data sets that cannot be modeled accurately, and students are asked to articulate why. Students use skills learned in Lesson 14 of Module 2 (where they used calculators to write linear regressions) and apply similar techniques for data sets that are better suited to modeling with quadratic or exponential regressions. Students use that same technique to find linear regressions and use their graphing calculators to examine the correlation coefficient and to find quadratic and exponential regressions. They compare correlation coefficients to determine which model is best for the data. Ultimately, students choose the regression model (linear, quadratic, or exponential) most appropriate to a given data set and then write, verify, and interpret these models in context. Students will need a graphing calculator to complete this lesson.

Refer to the following full modeling cycle during this lesson (Found on page 61 of the CCLS and page 72 of the CCSS).



Scaffolding:

- Students will be more engaged when working with relevant and real data that interests them. Websites that provide data sets are a good resource for classroom investigations.
- This lesson might need to be divided into two days if students need more time to master the technology.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Classwork

Opening Exercise (5 minutes)

Display the following data set, which cannot be modeled precisely. Ask the first question to start a class discussion.

Opening Exercise

What is this data table telling us?

Age (Years)	NYC Marathon Running Time (Minutes)
15	300
25	190
35	180
45	200
55	225
65	280

The relationship between the age of runners in the NYC Marathon and their running time.

In answering the question above, students may give more detail about the structure of the data and the relationships between the numbers. Let them brainstorm, but guide them to use the titles of the columns to inform them and not to read between the lines too much.

After an initial discussion of the table, read the questions aloud and have students independently write their answers to the following questions:

- What function type appears to be best suited to modeling this data? Why do you think so?
  - *It looks like this data might be modeled by a quadratic function or absolute value function because it decreases, reaches a minimum, and then increases again.*
- Can we model this data precisely using the methods learned in previous lessons? Why or why not?
  - *These specific data points cannot be modeled precisely by a linear function or absolute value function because even though the x-values are given at regular intervals, there is not a common first difference (on either side of the vertex) in the function values. Nor can the data be modeled by a quadratic function because the second differences are not exactly the same. An exponential function will not work either because there is not a common ratio between consecutive terms.*
- What questions might we want to ask, using the given data and its model(s)?
  - *Let students brainstorm the possibilities. Here are a few possibilities: How was the data collected? How many total data points are there? Are these running times averages? What is a realistic domain? Is it reasonable to include ages less than 15? How about ages greater than 65? Note: Later in this lesson, we look at the data and ask what the peak age is for running the marathon.*

To highlight MP.1 and MP.4, consider providing graph paper at this point and asking students to informally generate their own functions to model the data. These can then be compared with those found by the calculator.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Example 1 (15 minutes)

Remind students that most real-world data is messy and imperfect, like the examples they worked with in Module 2. Frequently, statisticians use technology to find the function model (linear, quadratic, etc.) that *best* fits the data, even if the fit is not perfect. In this lesson, students analyze regressions by considering the correlation coefficient and use regression models to answer questions and make predictions about the data. Remind students of the procedure they used in Lesson 14 of Module 2 to write linear regressions.

Example 1

Remember that in Module 2, we used a graphing display calculator (GDC) to find a linear regression model. If a linear model is not appropriate for a collection of data, it may be possible that a quadratic or exponential model will be a better fit. Your graphing calculator is capable of determining various types of regressions. Use a graphing display calculator (GDC) to determine if a data set has a better fit with a quadratic or exponential function. You may need to review entering the data into the stats application of your GDC.

When you are ready to begin, return to the data presented in the Opening Exercise. Use your graphing calculator to determine the function that best fits the data. Then, answer some questions your teacher will ask about the data.

Scaffolding:

Some students may need to review this calculator procedure from Module 2 before extending it to include quadratic or exponential regressions. If your students use a different type of calculator, you will need to modify these instructions:

Finding the Regression Line (TI-84 Plus)

- Step 1: From your home screen, press STAT.
- Step 2: From the STAT menu, select the EDIT option. (EDIT enter)
- Step 3: Enter the x-values of the data set in L1.
- Step 4: Enter the y-values of the data set in L2.
- Step 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. Most students should be familiar with option 4 or the slope-intercept form. Option 8 is essentially the same representation using a different letter to represent slope and y-intercept. Option 8 is the preferred option in statistical studies.)
- Step 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 function of 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.

First, it is important to explain that the quantity r, called the correlation coefficient, measures the strength and the

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

direction of a linear relationship between two variables. If the data appears to be quadratic or exponential, then your calculator may also show you the coefficient of determination, r^2. We can use the correlation coefficient for nonlinear relationships to determine which is the best fit, but we should never rely solely on the statistical correlation for nonlinear data. It is always important to also look at the scatter plot of the data along with the graph of the regression equation.

Finding a Regression Equation (TI-84 Plus)

Start with the same steps that you used for a linear regression.

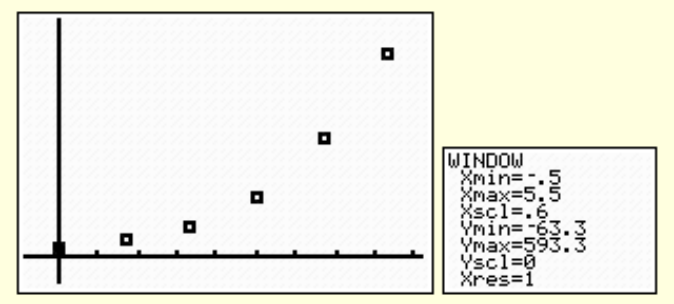
Step 1: From your home screen, press STAT.

Step 2: From the STAT menu, select the EDIT option. (EDIT enter)

Step 3: Enter the x-values of the data set in L1.

Step 4: Enter the y-values of the data set in L2.

Hours since observation began	Number of bacteria present
0	20
1	40
2	75
3	150
4	297
5	510



The example below shows the number of bacteria cells growing in a laboratory.

**After Step 4, however, the procedure changes. Instead of choosing option 4: LinReg$(ax + b)$, students will choose option 5: QuadReg, or option 0: ExpReg, depending on which type of function appears to be most likely.*

Step 5: Press [STAT]. Select menu item CALC and then select option 0: ExpReg. Press [ENTER].

Step 6: With ExpReg on the home screen, press [VARS] and select Y-VARS, FUNCTION, Y1 and press [ENTER] so that ExpReg Y1 is displayed on the home screen. Select [ENTER] to see results. The exponential regression will be stored in Y1.

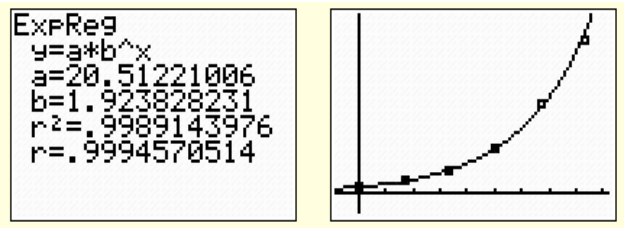
Step 7: To graph both the regression and the scatter plot of data, fit the window to match the parameters of the data set, and also to make sure that PLOT1 and Y1 are highlighted on the [Y=] screen.

Note: To find the coefficient of determination r^2, search the CATALOG by pressing [2ND][0], selecting DIAGNOSTIC ON, and pressing [ENTER][ENTER]. Once the diagnostic is turned on, every time students find a regression, the coefficient of determination r^2 will be listed at the bottom of the screen.

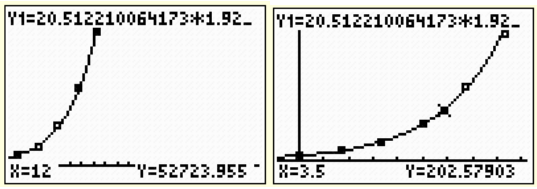
© 2014 Common Core, Inc. Some rights reserved. commoncore.org
BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

In the example below, we can see that the exponential function that closely models this data is approximately:

$f(x) = (20.51)(1.92)^x$ and that the correlation coefficient is $0.9989\ldots$



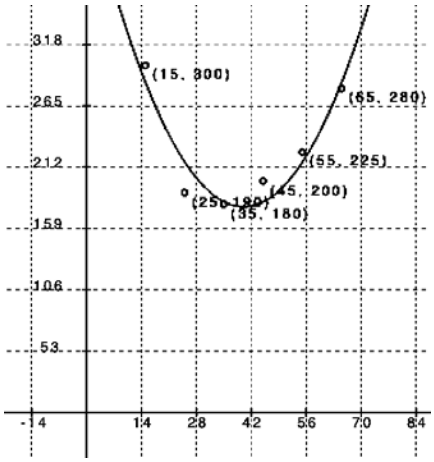
Also remember that once the function is graphed in your calculator, you can use the TRACE function to find specific values.

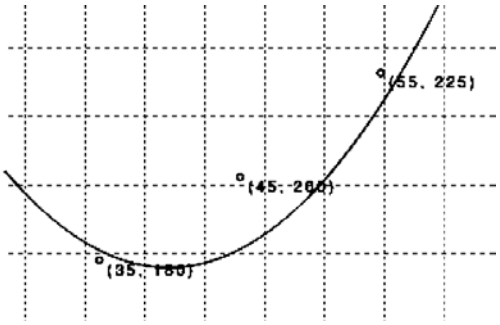


Now we return to the data we looked at in the Opening Exercise.

- We learned in Module 2 that a correlation coefficient close to 1 or -1 indicates that a regression line fits the data closely; the closer that r^2 is to 0, the less effective the regression will be at modeling the data and making predictions about the future. Have students use their graphing calculators to analyze the data from the Opening Exercise.
- Using the graphing calculator STAT function, we find that a linear regression has a correlation coefficient of 0.027, so we would not expect a linear function to model this data well.
- The correlation coefficient for an exponential regression is 0.068. This is also not expected to be a good fit.
- The quadratic regression model for this example is $f(x) = 0.172x^2 - 13.714x + 451.763$, which has a correlation coefficient of approximately -0.94.

Here is the graph of the data and the regression equation from two different views. In the first, we can see that the function is a good fit for the data we have. In the second, we zoom in on a few points and realize that, although the fit is not perfect, it is still very good.





- With how much accuracy do we need to approximate our results for this modeling function? For instance, if our leading coefficient is 0.1723214286, does that mean we have to round our answers in minutes to the ten-billionth decimal place?
  - *Typically, we can round to the same place value that our data was given in. Because our data set is originally given in terms of whole minutes, it is perfectly reasonable to round our answers to whole minutes as well.*

MP.6

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Keep in mind that rounding your answers to the nearest whole number does not mean you can round your coefficients to the same place value; in this example, that would mean rounding our a-coefficient from $0.17\ldots$ to 0, eliminating the quadratic term. For the regression coefficients, round to at least the hundredths or thousandths place, and further if the number is even smaller, or risk diminishing the accuracy of the model.

- Our model lies entirely in the first quadrant, and the model function has no (real) roots. Why is this?
  - *It would be impossible to run the marathon in 0 minutes or less, and it is impossible for age to be represented by a negative number.*
- According to this data, what is the peak age for running the marathon? In other words, what is the approximate best age for the shortest run time of this modeling function? What does this represent in the context of this problem? Do you find this data to be unusual in any way? If so, how?
  - *The vertex of our modeling function is approximately $(40, 179)$, meaning that the lowest possible marathon time of 179 minutes would be expected to be run by a 40-year-old, according to the model. This data could be interpreted as being somewhat strange because one might argue that the average 40-year-old is not as physically fit as the average 25- or 30-year-old.*
- Based on this regression model, how long might it take the average 50-year-old to run the NYC marathon?
  - *By substituting $x = 50$ into our regression model, we get $f(50) = 197$ minutes.*

Exercises (18 minutes)

Have students work with a partner or small group to answer the questions about the data set provided below. Then use the questions that follow to guide a class discussion in which you might call on students to share their answers.

Exercises

1. Use the following data table to construct a regression model, and then answer the questions.

Chicken Breast Frying Time (Minutes)	Moisture Content (%)
5	16.3
10	9.7
15	8.1
20	4.2
25	3.4
30	2.9
45	1.9
60	1.3

Data Source: ***Journal of Food Processing and Preservation*, 1995**

a. What function type appears to be the best fit for this data? Explain how you know.

The relationship between frying time and moisture content is best modeled by an exponential regression. Using the calculator yields the function $f(x) = 13.895(0.957)^x$ with a coefficient of determination of approximately 0.904. ($r = -0.95077\ldots$, so $R^2 = 0.9039\ldots$).

b. A student chooses a quadratic regression to model this data. Is he right or wrong? Why or why not?

This data cannot be modeled by quadratic regression because as cooking time increases, moisture content will always decrease and never begin to increase again. Also, in looking at the longer-term trend, we see that for a quadratic model the values are decreasing initially but will eventually begin to increase. This makes the quadratic model less reliable for larger x-values.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

c. **Will the moisture content for this product ever reach 0%? Why or why not?**

The moisture content will never reach 0% because exponential decay functions get smaller and smaller but never disappear entirely.

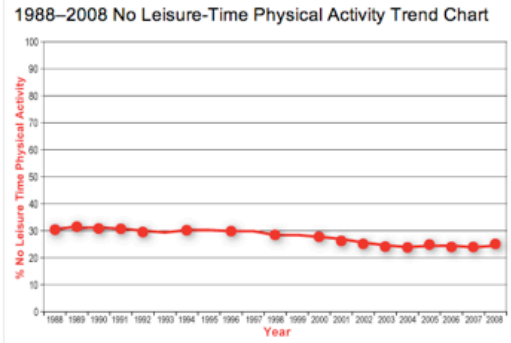
d. **Based on this model, what would you expect the moisture content to be of a chicken breast fried for 50 minutes?**

$f(50) = 13.895(0.957^{50}) = 1.5\%$ *moisture content for a chicken breast fried for 50 minutes.*

2. **Use the following data table to construct a regression model, then answer the questions based on your model.**

Prevalence of No Leisure-Time Activities, 1988 - 2008

Year	Years since 1988	% of prevalence
1988	0	30.5
1989	1	31.5
1990	2	30.9
1991	3	30.6
1992	4	29.3
1994	6	30.2
1996	8	28.4
1998	10	28.4
2000	12	27.8
2001	13	26.2
2002	14	25.1
2003	15	24.2
2004	16	23.7
2005	17	25.1
2006	18	23.9
2007	19	23.9
2008	20	25.1



Using technology to find a linear regression model, we find that the best-fit line is $y = -0.3988x + 31.517$, with a correlation coefficient of -0.952.

a. **What trends do you see in this collection of data?**

The data seem to be dropping gradually over the years but at a fairly constant, though small, negative rate. The correlation coefficient of nearly -1 indicates that this model has a strong negative linear relationship.

b. **How do you interpret this trend?**

The rate of leisure-time physical activity in the U.S. has slowly declined over the years since 1988 and is likely to continue to do so.

c. **If the trend continues, what would we expect the percentage of people in the U.S. who report no leisure-time physical activity to be in 2020?**

The year 2020 is year 32, since 1988, so $f(32) = -0.3988(32) + 31.517 = 18.755$. So, if this trend continues, we would expect about 19% of the population to report no leisure-time physical activity in 2020.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Closing (2 minutes)

- Data plots and other visual displays of data can help us determine the function type that appears to be the best fit for the data.
- When faced with messy, real-world data sets, it is relatively easy to use technology to find the best possible fit for a function to model the data.
- We can also experiment with transforming a parent function to manually create a model.

Lesson Summary

- **Using data plots and other visual displays of data, the function type that appears to be the best fit for the data can be determined. Using the correlation coefficient, the measure of the strength and the direction of a linear relationship can be determined.**
- **A graphing calculator can be used if the data sets are imperfect. To find a regression equation, the same steps will be performed as for a linear regression.**

Exit Ticket (5 minutes)

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Name _____ Date _____

Lesson 7: Modeling a Context from Data

Exit Ticket

Use the following data table to construct a regression model, and then answer the questions.

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

a. What is the best regression model for the data?

b. Based on your regression model, what height would you expect a person with a shoe length of 13.4 inches to be?

c. Interpret the value of your correlation coefficient in the context of the problem.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Exit Ticket Sample Solutions

Use the following data table to construct a regression model, and then answer the questions:

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

a. **What is the best regression model for the data?**

The best model for regression here is linear, modeled using the calculator as $f(x) = 3.657x + 25.277$ with a correlation coefficient of 0.6547.

b. **Based on your regression model, what height would you expect a person with a shoe length of 13.4 inches to be?**

$f(13.4) = 74 \rightarrow$ a person with shoes 13.4 inches long might be 74 inches tall.

c. **Interpret the value of your correlation coefficient in the context of the problem.**

Based on the correlation coefficient, there is a moderate positive linear relationship between shoe length and height.

Problem Set Sample Solutions

1. **Use the following data tables to write a regression model, and then answer the questions:**

Prescription Drug Sales in the United States Since 1995

Years Since 1995	Prescription Drug Sales (billions of USD)
0	68.6
2	81.9
3	103.0
4	121.7
5	140.7

a. **What is the best model for this data?**

The best model for this data would be an exponential regression, given by the function $f(t) = 65.736(1.161)^t$ with a correlation coefficient of 0.987 and a correlation of determination of 0.975.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

b. Based on your model, what were prescription drug sales in 2002? 2005?

2002: $f(7) = 186.9$, and 2005: $f(10) = 292.5$

c. For this model, would it make sense to input negative values for t into your regression? Why or why not?

Because there were prescription drug sales in the years prior to 1995, it would make sense to use negative numbers with this model (unless some drastic change in drug sales in 1995 makes this model inaccurate for preceding years).

2. Use the data below to answer the questions that follow:

Per Capita Ready-to-Eat Cereal Consumption in the United States per Year Since 1980

Years Since 1980	Cereal Consumption (lb.)
0	12
1	12
2	11.9
3	12.2
4	12.5
5	12.8
6	13.1
7	13.3
8	14.2
9	14.9

Years Since 1980	Cereal Consumption (lb.)
10	15.4
11	16.1
12	16.6
13	17.3
14	17.4
15	17.1
16	16.6
17	16.3
18	15.6
19	15.5

a. What is the best model for this data?

The best regression fit here is the quadratic $f(t) = -0.018t^2 + 0.637t + 10.797$ with correlation coefficient of 0.92 and a coefficient of determination (R^2) of 0.85.

b. Based on your model, what would you expect per capita cereal consumption to be in 2002? 2005?

According to the model, $f(22) = 16.1$ lb. of cereal, and $f(25) = 15.5$ lb.

(Note: Because this model has a little lower coefficient of determination (0.85), these predictions may not seem to fit well with the given data table.)

c. For this model, will it make sense to input t-values that return negative $f(t)$-values into your regression? Why or why not?

No, $f(t)$ values for this model would correspond to negative pounds of cereal consumed, which is impossible. Therefore, this model would only be useful over the domain where $f(t)$ is positive.

© 2014 Common Core, Inc. Some rights reserved. commoncore.org
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.