Scatterplots and Regression

Strand: Algebra and Functions

Topic: An introduction to scatterplots and regression

Primary SOL: AFDA.3 The student will collect and analyze data, determine the equation of the curve of best fit in order to make predictions, and solve practical problems using models of linear, quadratic, and exponential functions.

Related SOL: AFDA.1, AFDA.2, AFDA.4

Materials

- How Long Will I Live? activity sheet (attached)
- Graphing utility

Vocabulary

coefficient of determination (r^2 *), correlation coefficient (r), curve of best fit, exponential, extrapolation, interpolation, linear, quadratic, regression, scatterplot*

Student/Teacher Actions: What should students be doing? What should teachers be doing?
Time: 90 minutes

1. Ask, "What do you think is the average life expectancy for someone born today and living in the United States all of his or her life?" Record guesses and then ask, "How could we refine our prediction? Is there a way we can mathematically study this?"
2. Distribute the How Long Will I Live? activity sheet.
3. Have students enter the data into a graphing utility and sketch the graph of the scatterplot using the data. Ask students to revise their predictions based on their analysis of the scatterplot.
4. Ask, "Given the scatterplot, what function might best model the data?"
5. Have students calculate each regression model equation, find the correlation coefficients, and then graph each regression equation. Students should compare the equations and graphs to discuss the similarities and differences. All of the models may be similar in the window representing the given data, depending on the graphing utility being used.
6. Students should expand the window for the data to examine the "future" of each model. Suggested window: X[1930,2100] Y[60,100]

   Discuss the general pattern for each function and how that might apply to the future of the data. The quadratic model goes back down and probably isn't the best model for the situation. (Discuss with students what might happen in the future to cause the data to become more quadratic.) The exponential function increases rapidly, which seems unlikely even with medical advancements. The linear function continues to increase but at a slower rate; this would seem more realistic over the future. The logarithmic function will increase, but the rate of increase will slow over time and seems the most reasonable.

7. Have students look at the correlation coefficient and the coefficient of determination for each of the models and discuss what they notice about them. Remind students of the meaning of these coefficients.

   r (correlation coefficient) represents the strength and direction between two variables.

   r^2 (coefficient of determination) represents the percent of variance in a variable that can be explained by the relationship between the two variables.

   Students should be able to write a statement showing that they understand what r and r^2 mean in terms of the data. For example, if $r = 0.8$, then the data could indicate there is a moderate to strong relationship between the year a person is born and their life expectancy. That r value produces $r^2 = 0.64$, which indicates that 64 percent of the variance in the year a person is born can explain life expectancy.

   Students should notice that the correlation coefficients are close together but that the logarithmic function has the strongest coefficient (closer to 1). The quadratic has the strongest r^2 value, which could indicate that model as a good predictor for those values within the data set. Students should learn that it is important to look at the situation as well as the data when deciding on a model equation and/or using that model to make predictions.
8. Have students predict their life expectancy using their model equation; predict when the life expectancy would be 90 years. Discuss interpolation, extrapolation, and the likelihood that your predictions are accurate.

Assessment

- **Questions**
  - Explain which model best fits the life expectancy data and justify your choice in two ways.
  - Describe a practical situation in which a quadratic might fit, and explain why that model would best represent that situation.
- **Journal/writing prompts**
  - Compare the models for life expectancy both graphically and algebraically.
  - Discuss which model might fit the data the best when predicting within the domain of the given data vs. which model is better using the data to make future predictions.
  - A classmate insists that the quadratic is the best model for this data because the r^2 is higher. Write an argument to convince them that one of the other models is better over the long run. (Students' understanding of r^2 in terms of strength of the relationship between variables, predictability within intervals and overall, and application in specific context will affect their ability to address the argument.
- **Other Assessments**
  - Monitor student progress while they work in groups to find the regression equations and provide feedback.

  - Ask students to record their predictions on individual whiteboards to share with the class.

Extensions and Connections

- Have students research other practical situations that could represent the other models, particularly the exponential and quadratic functions.

Strategies for Differentiation

- Use vocabulary cards for related vocabulary listed above.
- Group work/jigsaw: Divide the class into four groups and have each group calculate one of the model equations. Have each group predict what their life expectancy would be based on their model equation. Reorganize groups so that each group has at least one person with each of the regression equations. Have the students compare the equations and predictions and discuss which model might work best.
- Give students a smaller data set.

Note: The following pages are intended for classroom use for students as a visual aid to learning.

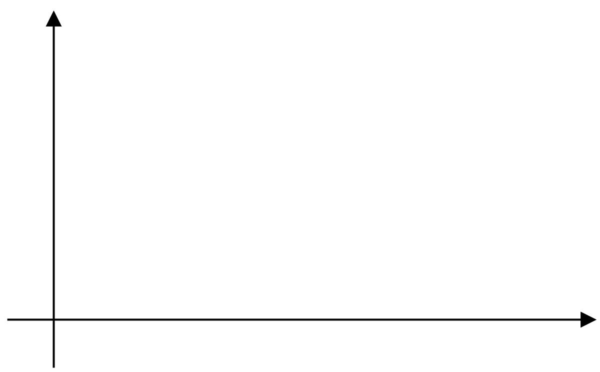
How Long Will I Live?

Below is data on life expectancy for all adults based on the year they were born. So, if you were born in 1940, you had an average life expectancy of 62.9 years.

Year	Life Expectancy	Year	Life Expectancy
1940	62.9	1990	75.4
1950	68.2	1995	75.8
1960	69.7	2000	76.8
1970	70.8	2005	77.6
1975	72.6	2010	78.7
1980	73.7	2014	78.8
1985	74.7		

Source: https://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_04.pdf (Table 8)

1. Use a graphing utility to enter and graph the data above. Sketch the scatterplot below.



2. Calculate the linear, quadratic, and exponential regression equations for the data. Round all coefficient values in the equations to the nearest hundredth. For each regression, record the correlation coefficient, *r*, rounded to the nearest ten-thousandth.
   a. Linear Regression Equation: ________________________________

      $r =$ ____________ Correlation: ____________________________________

      $r^2 =$ ___________ Correlation: ____________________________________

   b. Quadratic Regression Equation: _____________________________

      $r =$ ____________ Correlation: ____________________________________

      $r^2 =$ ___________ Correlation: ____________________________________

c. Exponential Regression Equation: ____

$r =$ ____________ Correlation: ______________________________________

$r^2 =$ ___________ Correlation: ______________________________________

d. Logarithmic Regression Equation: ____

$r =$ ____________ Correlation: ______________________________________

$r^2 =$ ___________ Correlation: ______________________________________

3. Graph the three models together with scatterplots using your graphing utility.
   a. What observations can you make about each model?

   b. What similarities do you notice about the graphs of the functions?

   c. What difference do you notice about the graphs of the functions?

   d. Which model seems to best fit the data?

4. We want to be sure our model fits practical situations over an extended period of time. Change your x-axis so that you can look further into the future, to observe the overall and end behavior.
   a. What do you notice about the quadratic function?

   b. What do you notice about the values of the correlation coefficients and how the graph fits the data?

   c. Thinking about the correlation coefficients, the graphs and the situation, which function do you think would best fit the data?

Interpolation: finding a value for a data point within the given data range **Extrapolation:** finding a value for a data point outside of the given data range.

5. Use your best fit equation to complete each of the following.
   a. What should your life expectancy be, on average?

   b. Based on the best fit equation, will the life expectancy ever be 90 years?