1

Like a Glove

Least Square Regressions

MATERIALS

Uncooked spaghetti
Graphing technology

Lesson Overview

Students informally determine a line of best fit by visual approximation of a hand-drawn line. They are then introduced to a formal method to determine the linear regression line of a data set using graphing technology; the mathematics behind the calculator function is explained using the related terms *Least Squares Method, regression line,* and *centroid*. Students then use the line of best fit to make predictions and distinguish between the terms *interpolation* and *extrapolation*.

Algebra 1

Linear Functions, Equations, and Inequalities

(3) The student applies the mathematical process standards when using graphs of linear functions, key features, and related transformations to represent in multiple ways and solve, with and without technology, equations, inequalities, and systems of equations. The student is expected to:

(C) graph linear functions on the coordinate plane and identify key features, including x-intercept, y-intercept, zeros, and slope, in mathematical and real-world problems.

(4) The student applies the mathematical process standards to formulate statistical relationships and evaluate their reasonableness based on real-world data. The student is expected to:

(C) write, with and without technology, linear functions that provide a reasonable fit to data to estimate solutions and make predictions for real-world problems.

Number and Algebraic Methods

(12) The student applies the mathematical process standards and algebraic methods to write, solve, analyze, and evaluate equations, relations, and functions. The student is expected to:

(A) decide whether relations represented verbally, tabularly, graphically, and symbolically define a function.

ELPS

1.A, 1.C, 1.E, 1.F, 1.G, 2.C, 2.E, 2.I, 3.D, 3.E, 4.B, 4.C, 5.B, 5.F, 5.G

Essential Ideas

- Interpolation is the process of using a regression equation to make predictions within the data set.
- Extrapolation is the process of using a regression equation to make predictions beyond the data set.
- A least squares regression line is the line of best fit that minimizes the squares of the distances of the points from the line.

Lesson Structure and Pacing: 2 Days

Day 1

Engage

Getting Started: Frozen Yogurt. . .When It's Freezing?

Students analyze data by creating a scatter plot and using a piece of spaghetti to estimate the line of best fit. They readjust the line of best fit (piece of spaghetti) as each data point is added to the scatter plot. Students conclude that with only one data value, an infinite number of lines are possible, and with two data values, only one line is possible. Once a third non-collinear point is introduced, students must make judgments about the appropriate position for the piece of spaghetti for the line of best fit. Students then estimate a line of best fit for the entire data set and interpret its meaning in terms of the problem situation.

Develop

Activity 1.1: A Line Of Best Fit

Students informally determine a line of best fit by visual approximation of a hand-drawn line and use their equation to make predictions. They are introduced to a formal method to determine the linear regression line of a data set using graphing technology; the mathematics behind the calculator function is explained using the related terms *Least Squares Method, regression line,* and *centroid*. Students calculate the least squares regression via graphing technology and use the function to make predictions. Finally, they compare the two sets of predictions.

Day 2

Activity 1.2: Making Predictions

Students use graphing technology to generate a regression line and then interpret the contextual and mathematical meanings of each element of the equation.

Activity 1.3: Making Predictions Within And Outside a Data Set

Students analyze a data set and use technology to create a regression equation. They then make predictions with values that lie within the parameters of the given domain and some predictions from values that lie outside the range of the given domain. The terms *interpolation* and *extrapolation* are defined.

Demonstrate

Talk the Talk: Tell Me Ev-ery-thing

Students discuss the use and accuracy of the regression line for making predictions for data points outside of the domain of the given data set.

ENGAGE

Getting Started: Frozen Yogurt. . .When It's Freezing?

Facilitation Notes

In this activity, students analyze data by creating a scatter plot and using a piece of spaghetti to estimate the line of best fit. They readjust the line of best fit (piece of spaghetti) as each data point is added to the scatter plot. Students conclude that with only one data value, an infinite number of lines are possible, and with two data values, only one line is possible. Once the third non-collinear point is introduced, students must make judgments about the appropriate position for the piece of spaghetti for the line of best fit. Students then estimate a line of best fit for the entire data set and interpret its meaning in terms of the problem situation.

Ask a student to read the introduction aloud. Discuss the scenario and directions as a class before distributing a piece of spaghetti to each student.

Have students work with a partner or in a group to complete Questions 1 and 2. Share responses as a class.

Questions to ask

- How many different lines can be determined using a single point?
- How does the position of your piece of spaghetti going through a single point compare to the position of your classmates' piece of spaghetti?
- How many different lines can be determined using two distinct points?
- Are any two points always collinear? Is it possible for two points to be non-collinear?
- How many points do you need to determine a unique linear function?
- How does the position of your piece of spaghetti going through two distinct points compare to the position of your classmates' piece of spaghetti?
- Does the line appear to have a positive or negative slope? How do you know?
- What does the sign of the slope tell you about the problem situation?
- Can three non-collinear points be represented using the same linear function?
- How can you write a linear function for three or more points if they are not all collinear?

- How does the position of your piece of spaghetti going through three non-collinear points compare to the position of your classmates' piece of spaghetti? Four points? Five points?
- How can you use your linear model to make predictions?

Summary

Different relationships can exist when you only analyze parts of data sets. To understand and describe relationships in data, the entire data set must be considered.

Activity 1.1

A Line Of Best Fit

DEVELOP

Facilitation Notes

In this activity, students informally determine a line of best fit by visual approximation of a hand-drawn line and use their equation to make predictions. They are introduced to a formal method to determine the linear regression line of a data set using graphing technology; the mathematics behind the calculator function is explained using the related terms *Least Squares Method, regression line,* and *centroid*. Students calculate the least squares regression via graphing technology and use the function to make predictions. They compare the two sets of predictions.

Have students work with a partner or in a group to complete Questions 1 through 3. Share responses as a class.

Questions to ask

- What strategy did you use to determine where to place your prediction line?
- How many of the data points are actually on your prediction line?
- What information did you use to write a prediction equation?
- How could you determine if your prediction line is a good representation of the data? The best representation?
- Did your predictions make sense?
- Now that you know other students' methods, would you change your method? Why or why not?

Have students work with a partner or in a group to complete Questions 4 through 7. Share responses as a class.

Differentiation strategies

- To assist all students, model the graphing technology process for the class, so that you may respond to technology questions as they arise and help students make sense of the process.

- To extend the activity, in addition to using the graphing technology to determine the equation, take the time to go through the process of creating the scatter plot, graphing the equation, and accessing the table of values for the equation. This allows students to see an exact graphical answer and have access to a table of values.

Questions to ask

- Does the linear function obtained from the technology give better predictions than your prediction equation? Which model is more accurate?
- How did the predictions calculated from your equation compare to those using the line of best fit generated from the calculator?
- For every 1° that the temperature increases, the number of customers goes up by 0.687. How could this be restated in terms of the increase in number of degrees per each additional customer?
- What is another way to restate the rate using reasonable numbers?

Misconception

Students sometimes confuse the two tables in the graphing calculator, one with the real data entered to create the linear regression equation and one with the values generated from the linear regression equation. To clear the confusion, select points from each table and have students locate them on the graph.

Ask a student to read the information and definitions following Question 7 aloud and complete Question 8 as a class.

Have students work with a partner or in a group to complete Questions 9 and 10. Share responses as a class.

Misconceptions

- Students are sometimes uncertain when to use the term *line of best fit* and when to use the term *regression line*. *Line of best fit* is a general term that can be used whether the line was estimated by being hand-drawn or calculated using a regression process. *Regression line* can be used only if that process was applied to determine the line.
- Students sometimes think the line of best fit must pass through points on the scatter plot, especially the points containing the smallest and largest x-values. To counteract this thinking, note that this is not the case with the Frozen Yogurt Problem.

Questions to ask

- Explain the calculations involved in the Least Squares Method.
- What is the sum of the squares of Alysse's vertical distances from the line?
- What is the sum of the squares of Bonito's vertical distances from the line?

- Who has data points closer to their line? Does this make their line a better fit?
- What has a bigger effect on the least squares value: a bunch of points somewhat close to the same line or most points very close to the same line and one point much farther from the line? Use mathematics to support your answer.

Differentiation strategy

To extend the activity, have students create a graph modeling Alysse's and Bonito's thinking.

Summary

Graphing calculators use the Least Squares Method to determine the line of best fit. The line of best fit can be used to model data and predict the dependent value when given an independent value.

Activity 1.2

Making Predictions

Facilitation Notes

In this activity, students use graphing technology to generate a regression line and then interpret the contextual and mathematical meanings of each element of the equation.

Have students work with a partner or in a group to complete Questions 1 through 5. Share responses as a class.

Questions to ask

- What does the Span column represent?
- Why do you think spans were used?
- How do you calculate the range of the data set?
- Is each *x*-value mapped onto exactly one *y*-value?
- Can a linear function be used to model this data set?
- Do the data appear to have a positive or negative association?
- How many consecutive spans can be described as having a decrease in average global temperature?

Have students work with a partner or in a group to complete Questions 6 through 10. Share responses as a class.

Questions to ask

- Will any or all of the data points on a scatter plot lie exactly on the regression line?

- How does your least squares regression equation compare to your classmates' equations?
- Does your line of best fit have a positive or negative slope? Explain why this makes sense for this context.
- Why do you think it is called a line of best fit?
- How does your prediction compare to your classmates' prediction? What would account for a different prediction?

Misconception

Students may assume that using the least squares method resulting in a line of best fit models the problem situation at every point of the domain rather than an appropriate subset of the domain. Be sure to always ask the students if the prediction makes sense in the particular problem situation.

Summary

A line of best fit is a way to model linear trends in real-world data.

Activity 1.3

Making Predictions Within and Outside a Data Set

Facilitation Notes

In this activity, students analyze a data set and use technology to determine a linear regression equation. They then make predictions for values that lie within the given domain and outside of the given domain. The terms *interpolation* and *extrapolation* are defined.

Have students work with a partner or in a group to complete Questions 1 through 7. Share responses as a class.

Differentiation strategies

To scaffold support,

- Discuss a method for entering the data for the independent variable into the graphing technology. If needed, suggest they represent the start year in the data set, 2010, as 0 on their data list.
- Have students relate the terms *interpolation* and *extrapolation* with words that have the same prefixes that they already know, such as *interior* and *exterior*. Discuss how interpolation and extrapolation relate to the interior and exterior of the data points.

Questions to ask

- What ordered pair did you use to represent the data point?
- How is the year 2010 represented in your data set?
- What is the value of the slope in your regression equation? How does it relate to the problem situation?
- What is the value of the *y*-intercept in your regression equation? How does it relate to the problem situation?
- How can interpolation and extrapolation help you judge the accuracy of the line of best fit?
- How does the graph of the problem situation compare to the graph of the regression equation?
- Why is the graph of the problem situation a subset of the graph of this linear function?
- Does a negative *y*-value make sense in this problem situation?
- Does a *y*-value greater than 100 make sense in this problem situation?

Summary

Lines of best fit are most appropriately used as predictors within the bounds of the domain of the given data. Caution must be exercised when using a prediction equation to make prediction outside the boundaries of the original data set.

Talk the Talk: Tell Me Ev-ery-thing

DEMONSTRATE

Facilitation Notes

In this activity, students discuss the use and accuracy of the regression line for making predictions for data points outside of the domain of the given data set.

Have students work with a partner or in a group to complete Questions 1 and 2. Share responses as a class.

Questions to ask

- Why is it impossible for you to complete the least squares regression model by hand?
- Are interpolation values or extrapolation values included in the domain of a problem situation? Explain.

Summary

A linear regression can be used to make predictions. Predictions made by extrapolation will likely be less accurate than predictions made by interpolation.

NOTES

1

Like a Glove

Least Squares Regressions

Warm Up

Use the slope formula to write an equation for each.

1. A line that passes through the point (2, 5) with a slope of 5

2. A line that passes through the points (5, 2) and (3, 8)

Learning Goals

- Create a graph of data points with and without technology.
- Determine an equation for a line of best fit by visual approximation of a hand-drawn line.
- Determine a linear regression equation using technology.
- Make predictions about data using a linear regression equation.
- Explain the calculations involved in the Least Squares Method.
- Choose a level of accuracy appropriate when reporting quantities.

Key Terms

- Least Squares Method
- centroid
- regression line
- interpolation
- extrapolation

You have searched for patterns in graphs and sequences of numbers. How can you use what you know to identify patterns in sets of data?

Warm Up Answers

1. $y = 5x - 5$
2. $y = -3x + 17$

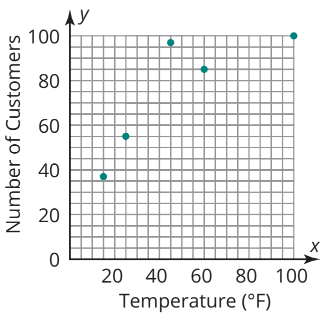
Answers

1a. Answers will vary.

1b. The data points can be connected to make a straight line.

1c. Answers will vary.

1d. Answers will vary.



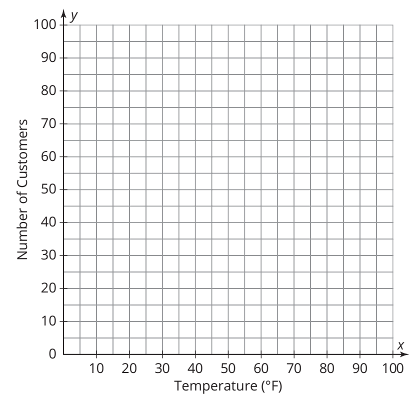
2. There is a positive association. As the temperature increases, the number of customers increases.

GETTING STARTED

Frozen Yogurt. . .When It's Freezing?

Mr. Templeton's Future Business Leaders Club (FBLC) is helping a frozen yogurt shop located near the school analyze how the business is affected by the weather. The owner is wondering whether there is a relationship between the temperature and the number of customers that buy yogurt during the 2 hours immediately after school. The FBLC collected this data.

Temperature (°F)	Number of Customers
45	97
25	55
60	85
15	37
100	100



1. Construct a scatter plot of the collected data.

a. Plot the first data point. Is there a pattern? Use a piece of spaghetti to approximate a line that models the data.

b. Add the second data point to the graph. Is there a pattern? Adjust the piece of spaghetti to approximate a line that models the data with the additional point.

c. Add the third data point to the graph. Describe the pattern that you see. Approximate the line using the spaghetti.

d. Continue this process until all five data points are plotted and recorded in the table.

2. Use your linear model to describe the relationship between the temperature outside and the number of customers at the frozen yogurt shop.

2 • TOPIC 3: Linear Regressions

ELL Tip

Make sure students understand what frozen yogurt is to make sense of the relationship between temperature and number of customers. Ask them to share examples of frozen treats from their culture.

ACTIVITY 1.1

A Line Of Best Fit

You have approximated the line that best represents the data with each additional data point.

1. Use the full data set and the line that you approximated to write an equation that you think best represents the data.

2. Based on your equation, predict the number of customers to visit the frozen yogurt shop in the two hours after school for each given temperature.

a. 85°F

b. 115°F

c. 10°F

3. Compare your predictions with your classmates. Did your predictions differ from the other groups? Explain why or why not.

Answers

1. Answers will vary.

2a-c. Answers will vary.

3. The predictions varied among my classmates because we wrote different equations for the line of best fit.

Answers

4. The temperature outside is the independent variable and the number of customers is the dependent variable. The independent variable represents the domain of the function, and the value of the dependent variable is generated by the value of the independent variable.
5. $y = 0.687x + 41.158$; the y-intercept means that the frozen yogurt shop could expect about 41 customers if the temperature were 0°F; the slope means that for every degree that the temperature increases, the number of customers goes up by 0.687 customers.

6a. At 85°F, the yogurt shop should expect approximately 99 customers.

6b. At 115°F, the yogurt shop should expect approximately 120 customers.

6c. At 10°F, the yogurt shop should expect approximately 48 customers.

7. Our predictions are the same because we are using the same linear regression line.

You have noticed that estimating a line of best fit can give different predictions. Fortunately, with technology you can create prediction equations as well as a scatter plots from tables of data. You just need to build a data table that has an independent variable and a dependent variable.

4. Identify the independent and dependent variable. What is the significance of those designations?

5. Use the data table and a graphing calculator to generate a line of best fit. What is the slope and *y*-intercept of the line and what do they represent?

6. Use the new line of best fit to predict the number of customers at the frozen yogurt shop immediately after school for each given temperature.

a. 85°F **b. 115°F**

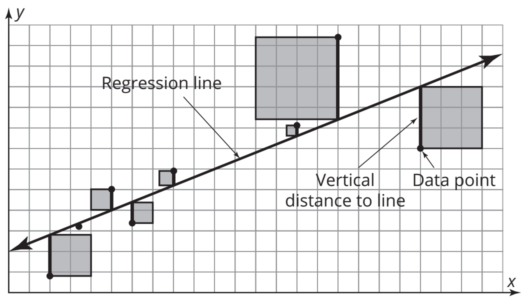
c. 10°F

7. How do your predictions compare to the predictions from the other groups?

4 • TOPIC 3: Linear Regressions

The equation that your calculator uses to give you the line of best fit is called the **Least Squares Method**. This is a method that creates a line of best fit for a scatter plot that has two basic requirements:

- The line must contain the *centroid* of the data set. The **centroid** is a point whose x-value is the mean of all the x-values of the points on the scatter plot and its y-value is the mean of all the y-values of the points on the scatter plot.
- Even though infinitely many lines can pass through the centroid, the **regression line** has the smallest possible vertical distances from each given data point to the regression line. The sum of the *squares* of those distances are at a minimum with this line.



8. Consider the graph of the sample regression line.

a. What do the vertical lines in bold represent?

b. What do the shaded squares represent? How do they relate to the Least Squares Method?

NOTES

Answers

8a. Each vertical line represents the difference in the y-value of a data point and the y-value of the point on the linear regression line for a given x.

8b. The area of each shaded square is equal to the vertical distance squared. The line that provides the minimum sum of the area of the squares is the least squares regression line.

Answers

9. Alysse is correct. The sum of the squares of the vertical distances from her line equals 20, which is lower than the sum of the squares of Bonito's, which is 40.
10. A linear regression line is a line that is fairly close to all data points instead of very close to most points, but one point is very distant from the line.

9. Alysse and Bonito each draw a regression line to model a set of data. They both record the vertical distances between each point and the regression line.

Alysse	**Bonito**
Vertical Distances: 2, 2, 2, 2, 2	Vertical Distances: 1, 1, 1, 1, 6

Both students believe they drew the least square regression line. Who's correct? Justify your choice.

10. How does your decision in Question 9 inform you about the placement of a line of best fit using the Least Squares Method?

ACTIVITY

1.2 Making Predictions

The table shown lists the average global temperature in 5-year spans from 1957 to 2016.

Years	Span	Average Temperature (°F)
1957–1961	1	57.250
1962–1966	2	57.121
1967–1971	3	57.196
1972–1976	4	57.189
1977–1981	5	57.495
1982–1986	6	57.445
1987–1991	7	57.780
1992–1996	8	57.700
1997–2001	9	58.053
2002–2006	10	58.262
2007–2011	11	58.244
2012–2016	12	58.448

1. What is the range of the data set?

2. Identify the independent and dependent variables and their units of measure.

3. Does the data represent a function? Does it appear that there is a specific function that could model this data set? If so, describe the function. If not, state why not.

4. Use technology to graph a scatter plot demonstrating the relationship between time spans and temperature. What association do you notice?

5. Between which consecutive spans was there a decrease in average global temperature?

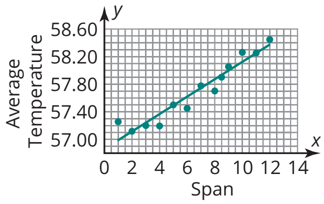
Answers

1. The range is 58.448 − 57.121 = 1.327°F.
2. The number of the span is the independent variable and the average temperature in °F is the dependent variable.
3. Each ordered pair has exactly one member of the domain paired with one member of the range, so it represents a function.

   It appears that the data is growing at a constant rate and could be modeled by a linear function.
4. There appears to be a positive association.
5. Between 1 and 2, 3 and 4, 5 and 6, 7 and 8, and 10 and 11.

Answers

6. $y = 0.1259x + 56.863$



7. There appears to be a positive association, which is reflected in the positive slope of the linear regression line. The line fits the data well.

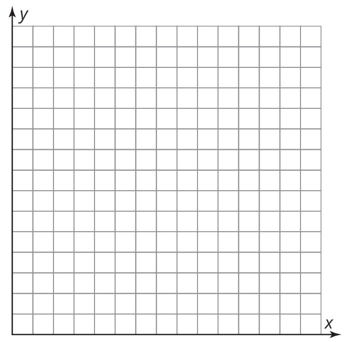
8. See table below.

9. 2032 – 2036 represents the 16th five-year span, so the prediction would be $y = 0.1259(16) + 56.863 = 58.877$°F.

ELL Tip

Students may be unfamiliar with the term *contextual meaning*. Discuss the definition of *context* and how *contextual meaning* is different from mathematical meaning. Ask students to demonstrate their understanding of the *contextual meaning* of the regression equation. If students provide different wording, discuss the differences and explain how the wording that explains the *contextual meaning* can be different as long as it correctly represents the context.

6. Use your graphing technology to determine the regression equation for the average global temperature data. Then sketch the data points and the line of best fit that you see.



7. What is the relationship between the equation for the line of best fit and any association you notice in the graph? Do you think that this line fits the data well?

8. For each expression from your linear regression equation about global temperatures, write an appropriate unit of measure and describe the contextual meaning. Then, choose a term from the word box to describe the mathematical meaning of each part.

Word Box
- input value
- output value
- rate of change
- *y*-intercept

		What it Means	
Expression	**Unit**	**Contextual Meaning**	**Mathematical Meaning**
$f(x)$			
0.1259			
x			
56.863			

9. Use your linear regression equation to predict the average global temperature for the years 2032–2036.

8 • TOPIC 3: Linear Regressions

		What it Means	
Expression	**Unit**	**Contextual Meaning**	**Mathematical Meaning**
$f(x)$	Degrees Fahrenheit	The predicted average temperature	Output value
0.1259	Degree rise per span of years	The predicted change in the average temperature per span number	Rate of change
x	Span number	Span number	Input value
56.863	Degrees Fahrenheit	The predicted average temperature for Span 0, Years 1952–1956	*y*-intercept

Answers

Answers may vary based on rounding.

1. $y = 5.6x + 4.9$ where x represents the number of years since 2010, and y represents the percent of revenue from streaming.
2. The percentage of US music sales revenue from streaming rises about 5.6% every year since 2010, when it was about 4.9%.
3. 21.7%; the predicted value percent is 0.7% higher than the actual value.
4. 10.5%; they are 1.5% apart.
5. Sample answer. Probably, as long as the line fits the data well.

ACTIVITY **1.3**

Making Predictions Within and Outside a Data Set

The music industry is constantly changing how it delivers music to its listeners. The table shows the percent of total U.S. music sales revenues from streaming.

Year	2010	2011	2012	2013	2014	2015
Percent of Total U.S. Music Sales Revenue From Streaming	7	9	15	21	27	34

1. Use graphing technology to determine the linear regression equation for the data.

2. Interpret the equation of the line in terms of this problem situation.

Ask yourself:

What is an appropriate level of accuracy needed throughout this situation?

If there is a linear association between the independent and dependent variables of a data set, you can use a linear regression to make predictions within the data set. Using a linear regression to make predictions within the data set is called **interpolation**.

3. Use your equation to predict the percent of streaming revenues in 2013. Compare the predicted value percent in 2013 with the actual value.

4. Compute the predicted value percent for 2011 and compare it with the actual value.

5. Do you think a prediction made using interpolation will always be close to the actual value? Explain your reasoning.

ELL Tip

Students may not be familiar with the terms *industry*, *revenues*, and *streaming*. Discuss the meaning of these terms using vocabulary that students understand and by providing examples. Review the data table at the beginning of the problem and clarify any remaining misunderstandings about the context of the problem.

Answers

6a. $x = 30$; 172.9%

6b. $x = -6$; -28.7%

7. No. More than 100% of music revenues from streaming isn't possible, and a negative percent from streaming isn't possible either.

Answers

1. Sample answer. The regression line generated using the graphing calculator uses the Least Squares Method for all possible cases and includes the centroid.
2. Sample answer. The regression line best fits the data given. Data outside of the given set may vary widely due to changes in the situation, or it may not be reasonable for the data to grow the same way outside the given set.

To make predictions for values of x that are outside of the data set is called **extrapolation**.

6. Use the equation to predict the percent of streaming revenues:

a. in 2040. **b. in 2004.**

7. Are these predictions reasonable? Explain your reasoning.

NOTES

TALK the TALK

Tell Me Ev-ery-thing

You have used technology to determine linear regression equations. You have then used those linear regression equations to predict unknown values within and without a data set.

1. Why is the linear regression line generated using technology more accurate than the line of best fit that can be written using two points?

2. Why are predictions made by extrapolation more likely to be less accurate than predictions made by interpolation?