Lesson 11: Using Linear Models in a Data Context

Student Outcomes

- Students recognize and justify that a linear model can be used to fit data.
- Students interpret the slope of a linear model to answer questions or to solve a problem.

Lesson Notes

In a previous lesson, students were given bivariate numerical data where there was an exact linear relationship between two variables. Students identified which variable was the predictor variable (i.e., independent variable) and which was the predicted variable (i.e., dependent variable). They found the equation of the line that fit the data and interpreted the intercept and slope in words in the context of the problem. Students also calculated a prediction for a given value of the predictor variable. This lesson introduces students to data that are not exactly linear but that have a linear trend. Students informally fit a line and use it to make predictions and answer questions in context.

Although students may want to rely on using symbolic representations for lines, it is important to challenge them to express their equations in words in the context of the problem. Keep emphasizing the meaning of slope in context, and avoid the use of "rise over run." Slope is the impact that increasing the value of the predictor variable by one unit has on the predicted value.

Classwork

Exercise 1 (10–12 minutes)

Introduce the data in the exercise. Using a short video may help students (especially English language learners) to better understand the context of the data. Then, work through each part of the exercise as a class. Ask students the following:

- Looking at the table, what trend appears in the data?
  - *There is a positive trend. As one variable increases in value, so does the other.*
- Looking at the scatter plots, is there an exact linear relationship between the variables?
  - *No, the four points cannot be connected by a straight line.*

Exercises

1. **Old Faithful is a geyser in Yellowstone National Park. The following table offers some rough estimates of the length of an eruption (in minutes) and the amount of water (in gallons) in that eruption.**

Length (minutes)	1.5	2	3	4.5
Amount of Water (gallons)	3,700	4,100	6,450	8,400

This data is consistent with actual eruption and summary statistics that can be found at the following links:

http://geysertimes.org/geyser.php?id=OldFaithful and http://www.yellowstonepark.com/2011/07/about-old-faithful/

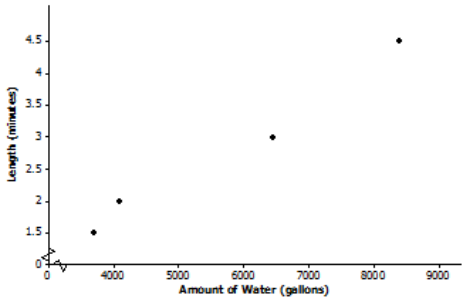
a. Chang wants to predict the amount of water in an eruption based on the length of the eruption. What should he use as the dependent variable? Why?

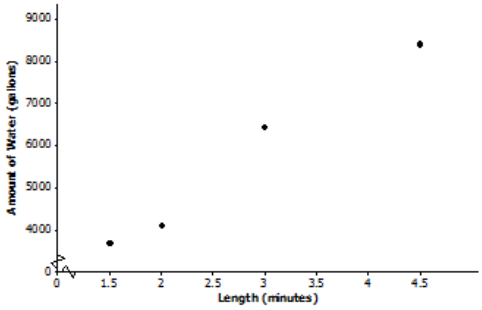
Since Chang wants to predict the amount of water in an eruption, the time length (in minutes) is the predictor, and the amount of water is the dependent variable.

Scaffolding:
Make the interchangeability of the terms *linearly related* and *linear relationship* clear to students.

b. Which of the following two scatter plots should Chang use to build his prediction model? Explain.

The predicted variable goes on the vertical axis with the predictor on the horizontal axis. So, the amount of water goes on the y-axis. The plot on the graph on the right should be used.





c. Suppose that Chang believes the variables to be linearly related. Use the *first* and *last* data points in the table to create a linear prediction model.

$$m = \frac{8400 - 3700}{4.5 - 1.5} \approx 1,566.7$$

So, $y = a + (1,566.7)x$.

Using either $(1.5, 3700)$ or $(4.5, 8400)$ allows students to solve for the intercept. For example, solving $3,700 = a + (1,566.7)(1.5)$ for a yields $a = 1,349.95$, or rounded to $1,350.0$ gallons. Be sure students talk through the units in each step of the calculations.

The (informal) linear prediction model is $y = 1,350.0 + 1,566.7x$. The amount of water (y) is in gallons, and the length of the eruption (x) is in minutes.

d. A friend of Chang's told him that Old Faithful produces about $3,000$ gallons of water for every minute that it erupts. Does the linear model from part (c) support what Chang's friend said? Explain.

This question requires students to interpret slope. An additional minute in eruption length results in a prediction of an additional $1,566.7$ gallons of water produced. So, Chang's friend who claims Old Faithful produces $3,000$ gallons of water a minute must be thinking of a different geyser.

e. Using the linear model from part (c), does it make sense to interpret the y-intercept in the context of this problem? Explain.

No, it doesn't make sense because if the length of an eruption is 0, then it cannot produce $1,350$ gallons of water. (Convey to students that some linear models have y-intercepts that do not make sense within the context of a problem.)

Exercise 2 (15–20 minutes)

Let students work in small groups or with a partner. Introduce the data in the table. Note that the mean times of the three medal winners are provided for each year. Let students work on the exercise, and confirm answers to parts (c)–(f) as a class. After answers have been confirmed, ask the class:

- What is the meaning of the y-intercept from part (c)?
  - *The y-intercept from part (c) is $(0, 34.91)$. It does not make sense within the context of the problem. In Year 0, the mean medal time was 34.91 seconds.*

2. The following table gives the times of the gold, silver, and bronze medal winners for the men's 100-meter race (in seconds) for the past 10 Olympic Games.

Year	2012	2008	2004	2000	1996	1992	1988	1984	1980	1976
Gold	9.63	9.69	9.85	9.87	9.84	9.96	9.92	9.99	10.25	10.06
Silver	9.75	9.89	9.86	9.99	9.89	10.02	9.97	10.19	10.25	10.07
Bronze	9.79	9.91	9.87	10.04	9.90	10.04	9.99	10.22	10.39	10.14
Mean Time	9.72	9.83	9.86	9.97	9.88	10.01	9.96	10.13	10.30	10.09

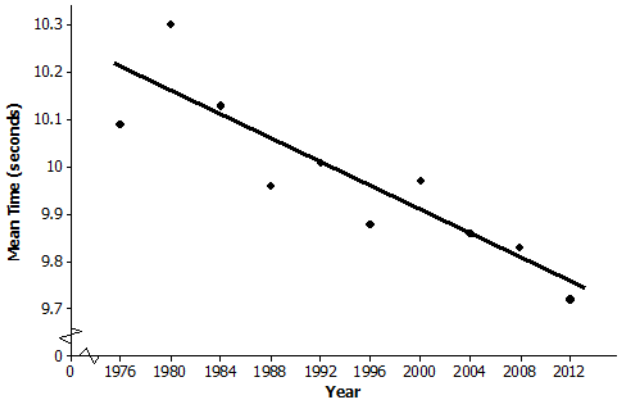
Data Source: https://en.wikipedia.org/wiki/100_metres_at_the_Olympics#Men

a. If you wanted to describe how mean times change over the years, which variable would you use as the independent variable, and which would you use as the dependent variable?

Mean medal time (dependent variable) is being predicted based on year (independent variable).

b. Draw a scatter plot to determine if the relationship between mean time and year appears to be linear. Comment on any trend or pattern that you see in the scatter plot.

The scatter plot indicates a negative trend, meaning that, in general, the mean race times have been decreasing over the years even though there is not a perfect linear pattern.



MP.7

c. **One reasonable line goes through the 1992 and 2004 data. Find the equation of that line.**

The slope of the line through $(1992, 10.01)$ *and* $(2004, 9.86)$ *is* $\frac{10.01 - 9.86}{1992 - 2004} = -0.0125$.

To find the intercept using $(1992, 10.01)$, *solve* $10.01 = a + (-0.0125)(1992)$ *for* a, *which yields* $a = 34.91$.

The equation that predicts the mean medal race time for an Olympic year is $y = 34.91 + (-0.0125)x$. *The mean medal race time* (y) *is in seconds, and the time* (x) *is in years.*

Note to Teacher: In Algebra I, students learn a formal method called least squares *for determining a "best-fitting" line. For comparison, the least squares prediction line is* $y = 34.3562 + (-0.0122)x$.

MP.2

d. **Before he saw these data, Chang guessed that the mean time of the three Olympic medal winners decreased by about** 0.05 **second from one Olympic Game to the next. Does the prediction model you found in part (c) support his guess? Explain.**

The slope -0.0125 *means that from one calendar year to the next, the predicted mean race time for the top three medals decreases by* 0.0125 *second. So, between successive Olympic Games, which occur every four years, the predicted mean race time is reduced by* 0.05 *second because* $4(0.0125) = 0.05$.

e. **If the trend continues, what mean race time would you predict for the gold, silver, and bronze medal winners in the 2016 Olympic Games? Explain how you got this prediction.**

If the linear pattern were to continue, the predicted mean time for the 2016 Olympics is 9.71 *seconds because* $34.91 - (0.0125)(2016) = 9.71$.

f. **The data point** $(1980, 10.3)$ **appears to have an unusually high value for the mean race time (**10.3**). Using your library or the Internet, see if you can find a possible explanation for why that might have happened.**

The mean race time in 1980 was an unusually high 10.3 *seconds. In their research of the 1980 Olympic Games, students find that the United States and several other countries boycotted the games, which were held in Moscow. Perhaps the field of runners was not the typical Olympic quality as a result. Atypical points in a set of data are called* outliers. *They may influence the analysis of the data.*

Following these two examples, ask students to summarize (in written or spoken form) how to make predictions from data.

Closing (2–3 minutes)

If time allows, revisit the linear model from Exercise 2. Explain that the data can be modified to create a model in which the y-intercept makes sense within the context of the problem.

Year	2012	2008	2004	2000	1996	1992	1988	1984	1980	1976
Number of Years (since 1976)	36	32	28	24	20	16	12	8	4	0
Gold	9.63	9.69	9.85	9.87	9.84	9.96	9.92	9.99	10.25	10.06
Silver	9.75	9.89	9.86	9.99	9.89	10.02	9.97	10.19	10.25	10.07
Bronze	9.79	9.91	9.87	10.04	9.90	10.04	9.99	10.22	10.39	10.14
Mean Time	9.72	9.83	9.86	9.97	9.88	10.01	9.96	10.13	10.30	10.09

Data Source: https://en.wikipedia.org/wiki/100_metres_at_the_Olympics#Men

- Using the data points for 1992 and 2004, $(16, 10.01)$ and $(28, 9.86)$, the linear model is $y = 10.21 + (-0.0125)x$.
- Note that the slope is the same as the linear model in Exercise 2.
- The y-intercept is now $(0, 10.21)$, which means that in 1976 (0 years since 1976), the mean medal time was 10.21 seconds.

Review the Lesson Summary with students.

Lesson Summary

In the real world, it is rare that two numerical variables are exactly linearly related. If the data are roughly linearly related, then a line can be drawn through the data. This line can then be used to make predictions and to answer questions. For now, the line is informally drawn, but in later grades more formal methods for determining a best-fitting line are presented.

Exit Ticket (8–10 minutes)

Name _____ Date _____

Lesson 11: Using Linear Models in a Data Context

Exit Ticket

According to the Bureau of Vital Statistics for the New York City Department of Health and Mental Hygiene, the life expectancy at birth (in years) for New York City babies is as follows.

Year of Birth	2001	2002	2003	2004	2005	2006	2007	2008	2009
Life Expectancy	77.9	78.2	78.5	79.0	79.2	79.7	80.1	80.2	80.6

Data Source: http://www.nyc.gov/html/om/pdf/2012/pr465-12_charts.pdf

a. If you are interested in predicting life expectancy for babies born in a given year, which variable is the independent variable, and which is the dependent variable?

b. Draw a scatter plot to determine if there appears to be a linear relationship between the year of birth and life expectancy.

c. Fit a line to the data. Show your work.

d. Based on the context of the problem, interpret in words the intercept and slope of the line you found in part (c).

e. Use your line to predict life expectancy for babies born in New York City in 2010.

Exit Ticket Sample Solutions

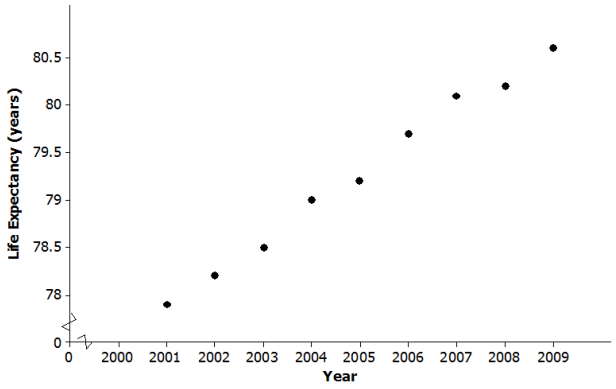
According to the Bureau of Vital Statistics for the New York City Department of Health and Mental Hygiene, the life expectancy at birth (in years) for New York City babies is as follows.

Year of Birth	2001	2002	2003	2004	2005	2006	2007	2008	2009
Life Expectancy	77.9	78.2	78.5	79.0	79.2	79.7	80.1	80.2	80.6

Data Source: http://www.nyc.gov/html/om/pdf/2012/pr465-12_charts.pdf

a. **If you are interested in predicting life expectancy for babies born in a given year, which variable is the independent variable, and which is the dependent variable?**

Year of birth is the independent variable, and life expectancy in years is the dependent variable.



b. **Draw a scatter plot to determine if there appears to be a linear relationship between the year of birth and life expectancy.**

Life expectancy and year of birth appear to be linearly related.

c. **Fit a line to the data. Show your work.**

Answers will vary. For example, the line through $(2001, 77.9)$ *and* $(2009, 80.6)$ *is* $y = -597.438 + (0.3375)x$*, where life expectancy (y) is in years, and the time (x) is in years.*

Note to Teacher: The formal least squares line (Algebra I) is $y = -612.458 + (0.345)x$*.*

d. **Based on the context of the problem, interpret in words the intercept and slope of the line you found in part (c).**

Answers will vary based on part (c). The intercept says that babies born in New York City in Year 0 *should expect to live around* -597 *years! Be sure students actually say that this is an unrealistic result and that interpreting the intercept is meaningless in this problem. Regarding the slope, for an increase of* 1 *in the year of birth, predicted life expectancy increases by* 0.3375 *year, which is a little over four months.*

e. **Use your line to predict life expectancy for babies born in New York City in 2010.**

Answers will vary based on part (c).

$-597.438 + (0.3375)(2010) = 80.9$

Using the line calculated in part (c), the predicted life expectancy for babies born in New York City in 2010 is 80.9 *years, which is also the value given on the website.*

Problem Set Sample Solutions

1. From the United States Bureau of Census website, the population sizes (in millions of people) in the United States for census years 1790–2010 are as follows.

Year	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880	1890
Population Size	3.9	5.3	7.2	9.6	12.9	17.1	23.2	31.4	38.6	50.2	63.0

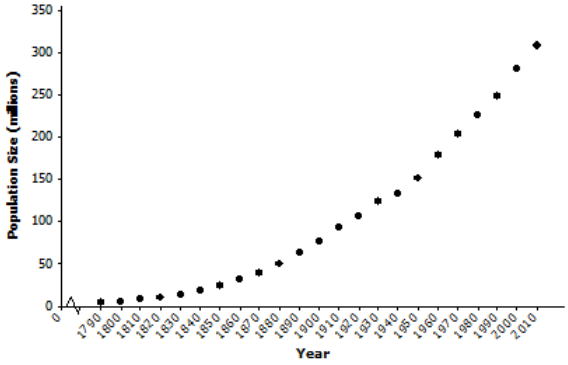
Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
Population Size	76.2	92.2	106.0	123.2	132.2	151.3	179.3	203.3	226.5	248.7	281.4	308.7

a. If you wanted to be able to predict population size in a given year, which variable would be the independent variable, and which would be the dependent variable?

Population size (dependent variable) is being predicted based on year (independent variable).

b. Draw a scatter plot. Does the relationship between year and population size appear to be linear?

The relationship between population size and year of birth is definitely nonlinear. Note that investigating nonlinear relationships is the topic of the next two lessons.



c. Consider the data only from 1950 to 2010. Does the relationship between year and population size for these years appear to be linear?

Drawing a scatter plot using the 1950–2010 data indicates that the relationship between population size and year of birth is approximately linear, although some students may say that there is a very slight curvature to the data.

d. One line that could be used to model the relationship between year and population size for the data from 1950 to 2010 is $y = -4875.021 + 2.578x$. Suppose that a sociologist believes that there will be negative consequences if population size in the United States increases by more than $2\frac{3}{4}$ million people annually. Should she be concerned? Explain your reasoning.

This problem is asking students to interpret the slope. Some students will no doubt say that the sociologist need not be concerned, since the slope of 2.578 million births per year is smaller than her threshold value of 2.75 million births per year. Other students may say that the sociologist should be concerned, since the difference between 2.578 and 2.75 is only $172,000$ births per year.

e. Assuming that the linear pattern continues, use the line given in part (d) to predict the size of the population in the United States in the next census.

The next census year is 2020.

$$-4875.021 + (2.578)(2020) = 332.539$$

The given line predicts that the population then will be 332.539 million people.

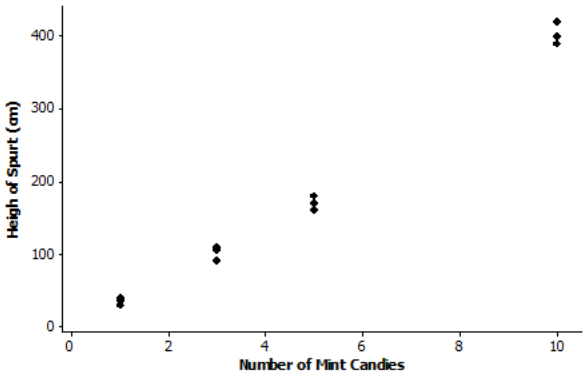
2. In search of a topic for his science class project, Bill saw an interesting YouTube video in which dropping mint candies into bottles of a soda pop caused the soda pop to spurt immediately from the bottle. He wondered if the height of the spurt was linearly related to the number of mint candies that were used. He collected data using 1, 3, 5, and 10 mint candies. Then, he used two-liter bottles of a diet soda and measured the height of the spurt in centimeters. He tried each quantity of mint candies three times. His data are in the following table.

Number of Mint Candies	1	1	1	3	3	3	5	5	5	10	10	10
Height of Spurt (centimeters)	40	35	30	110	105	90	170	160	180	400	390	420

a. Identify which variable is the independent variable and which is the dependent variable.

Height of spurt is the dependent variable, and number of mint candies is the independent variable because height of spurt is being predicted based on number of mint candies used.

b. Draw a scatter plot that could be used to determine whether the relationship between height of spurt and number of mint candies appears to be linear.



Scaffolding:

- The word *spurt* may need to be defined for English language learners.
- A spurt is a sudden stream of liquid or gas forced out under pressure. Showing a visual aid to accompany this exercise may help student comprehension.

c. Bill sees a slight curvature in the scatter plot, but he thinks that the relationship between the number of mint candies and the height of the spurt appears close enough to being linear, and he proceeds to draw a line. His eyeballed line goes through the mean of the three heights for three mint candies and the mean of the three heights for 10 candies. Bill calculates the equation of his eyeballed line to be

$$y = -27.617 + (43.095)x,$$

where the height of the spurt (y) in centimeters is based on the number of mint candies (x). Do you agree with this calculation? He rounded all of his calculations to three decimal places. Show your work.

Yes, Bill's equation is correct.

The slope of the line through $(3, 101.667)$ *and* $(10, 403.333)$ *is* $\frac{403.333 - 101.667}{10 - 3} = 43.095$.

The intercept could be found by solving $403.333 = a + (43.095)(10)$ *for* a, *which yields* $a = -27.617$.

So, a possible prediction line is $y = -27.617 + (43.095)x$.

d. In the context of this problem, interpret in words the slope and intercept for Bill's line. Does interpreting the intercept make sense in this context? Explain.

The slope is 43.095, *which means that for every mint candy dropped into the bottle of soda pop, the height of the spurt increases by* 43.095 cm.

The y*-intercept is* $(0, -27.617)$. *This means that if no mint candies are dropped into the bottle of soda pop, the height of the spurt is* -27.617 ft. *This does not make sense within the context of the problem.*

e. If the linear trend continues for greater numbers of mint candies, what do you predict the height of the spurt to be if 15 mint candies are used?

$$-27.617 + (43.095)(15) = 618.808$$

The predicted height would be 618.808 cm, *which is slightly over* 20 ft.