Lesson 2-2

# Linear Models

**BIG IDEA** The sum of squared deviations is a statistic for determining which of two lines fits the data better.

## Vocabulary

linear function
linear model
interpolation
extrapolation
observed values
predicted values
residual
sum of squared residuals

A **linear function** is a set of ordered pairs $(x, y)$ satisfying an equation of the form $y = mx + b$, where $m$ and $b$ are constants. Recall that the graph of every such function is a line with slope $m$ and $y$-intercept $b$.

**Mental Math**

What is the sum of a set 25 elements whose mean is 300?

## Fitting a Line to Data

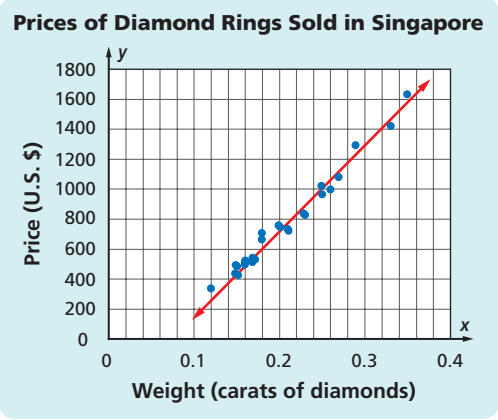
When data in a scatterplot lie near a line, we can create a **linear model** for the data, that is, a model of one variable as a linear function of the other. Even if the linear function does not contain all of the data points, it may still be useful in describing the overall trend of the data or in predicting values of either the dependent or independent variable.

In this lesson and the next, you will learn several techniques for constructing linear models. One technique is to fit a line to data "by eye," that is, to draw a line that seems close to the data.

**GUIDED**

### Example 1

Jewelers emphasize that the price of a diamond is determined by cut, carat weight, color and clarity. The table at the right gives carat weights and approximate prices in U.S. dollars for twenty diamond rings sold at a recent auction in Singapore. All rings are of the same quality gold and contain a single diamond. The data are graphed below.




| Weight $x$ | Price $y$ (U.S. $) |
|---|---|
| 0.18 | 702.00 |
| 0.17 | 517.50 |
| 0.25 | 963.00 |
| 0.29 | 1290.00 |
| 0.27 | 1080.00 |
| 0.15 | 484.50 |
| 0.20 | 747.00 |
| 0.25 | 1017.00 |
| 0.21 | 724.50 |
| 0.17 | 529.50 |
| 0.35 | 1629.00 |
| 0.33 | 1417.50 |
| 0.26 | 994.50 |
| 0.16 | 513.00 |
| 0.12 | 334.50 |
| 0.18 | 664.50 |
| 0.15 | 430.50 |
| 0.16 | 507.00 |
| 0.16 | 498.00 |
| 0.23 | 829.50 |

Source: Journal of Statistics Education

**The linear model is based on the weight of the diamond used. Although the data are not collinear, the line through (0.18, 600) and (0.32, 1400) seems close to the points. It has been added to the graph. An equation of this graphed line is one linear model for these data. Is the size alone a good predictor of price?**

a. **Find an equation of the graphed line which relates weight and price.**

b. **Interpret the slope of the line in the context of the problem.**

c. **Use the model to estimate the price of a 0.3-carat diamond ring.**

d. **Why is the model not good for predicting the cost of a 0.05-carat diamond ring?**

e. **Why is the set of data not a function?**

**Hard Rock**

**The largest rough diamond ever found weighed 3106 carats.**

**Solution**

a. We use the points (0.18, 600) and (0.32, 1400) to find an equation of the line:

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \underline{\ ?\ } \approx \underline{\ ?\ }$$

Substitute $m = 5714.29$ and (0.18, 600) into the point-slope form of the equation for a line.

$$\begin{aligned} y - y_1 &= m(x - x_1) \\ y - \underline{\ ?\ } &= 5714.29(x - \underline{\ ?\ }) \\ y &= \underline{\ ?\ } \end{aligned}$$

b. The slope is the rate of change. The cost increases by about ___?___ dollars for every 1-carat increase in weight.

c. To predict the cost from the weight, substitute $x = 0.3$ and solve for $y$.
When $x = 0.3$, $y =$ ___?___. According to the model, the cost of a 0.3-carat diamond ring will be about ___?___.

d. Substitute 0.05 for $x$ and solve for $y$. The model predicts that a 0.05-carat diamond ring will cost ___?___. This means that the seller would pay you to take the ring! This is not plausible, so the model is not a good predictor.

e. There is at least one $x$-value that does not have a unique $y$-value. There are two diamonds at 0.17 carats that have different prices, so the data set is not a function.

In this example, known diamond weights range from 0.12 to 0.35 carats. If you use the model to predict the price of a 0.3-carat diamond, you are making a prediction *between* known values. Prediction between known values is called **interpolation.** If you calculate the price of a diamond weighing 0.05 carats, you are making a prediction *beyond* known values. Prediction beyond known values is called **extrapolation**. Extrapolation is usually more hazardous than interpolation, because it depends on an assumption that a relationship will continue past the known data. In this case, a diamond much smaller than those in the sample might be inexpensive, but it will not be free.

## Measuring How Well a Line Models Data

In Chapter 1, you studied the sample variance, which was computed as the sum of the squared deviations from the mean divided by $n - 1$. A similar statistic, the *residual*, tells how far away data are from your chosen model. At the right is a table showing a smaller sample of the diamond ring data on page 87. These data, collected from sources such as experiments or surveys, are called **observed values**. Below, the scatterplot of these data is shown together with the graph of the line with equation $y = 2400x + 400$. This equation seems to be a pretty good model for the data. However, is it the best model? To compare models, we calculate *residuals*. The values predicted by a model are called **predicted values**. The observed value minus the predicted value is the **residual**. A residual is positive when the observed value is higher than what is predicted by the model. A residual is negative when the observed value is lower than what the model predicts.

**Diamond Ring Prices by Weight of Diamond**

| Weight | Price (U.S. dollars) |
|---|---|
| 0.15 | 484.50 |
| 0.16 | 507.00 |
| 0.18 | 702.00 |
| 0.25 | 963.00 |
| 0.27 | 1080.00 |
| 0.33 | 1417.50 |
| 0.23 | 829.50 |

For instance, a 0.16-carat diamond ring sold for \$507.00. This is lower than the price predicted by the model.

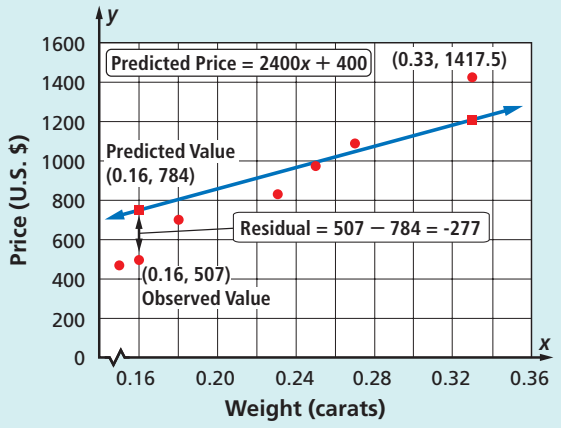
$$\begin{aligned} \text{predicted price} &= 2400 \cdot (0.16) + 400 \\ &= \$784 \end{aligned}$$

The residual, or error, is

$$\begin{aligned} \text{residual} &= \text{observed value} - \text{predicted value} \\ &= 507 - 784 \\ &= -277. \end{aligned}$$

The actual cost was \$277 less than predicted. On the graph, we see that the observed value is 277 units below the linear model $y = 2400x + 400$. That is why the residual is negative. Every data point has a residual, so $n$ data points provide $n$ residuals. The absolute value of a residual is the length of the vertical segment from the data point to the corresponding point on the linear model.

**Diamond Ring Prices by Weight of Diamond**




STOP QY

▶ **QY**

What is the residual for the 0.33-carat diamond?

On the graphs below, this line and another linear model are drawn to fit the seven points. Recall that when you compute variance, you add the squared deviations from the mean. Similarly, when modeling data with a line, we can measure the variation from the line by adding the squares of the residuals. You can use spreadsheets to calculate residuals and the sum of squared residuals, as shown below.

| | A weight | B price | C predicted | D residuals |
|---|---|---|---|---|
| ◆ | | | =2400*a[]+400 | =b[]−c[] |
| 1 | 0.15 | 484.5 | 760. | -275.5 |
| 2 | 0.16 | 507 | 784. | -277. |
| 3 | 0.18 | 702 | 832. | -130. |
| 4 | 0.25 | 963 | 1000. | -37. |
| 5 | 0.27 | 1080 | 1048. | 32. |

D residuals:=b[]−c[]

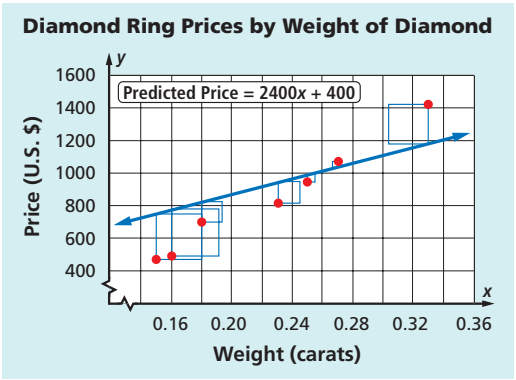
| | C predicted | D residuals | E sqresid | F sumsq... |
|---|---|---|---|---|
| ◆ | =2400*a[]+4 | =b[]−c[] | =d[]^2 | |
| 1 | 760. | -275.5 | 75900.3 | 237779. |
| 2 | 784. | -277. | 76729. | |
| 3 | 832. | -130. | 16900. | |
| 4 | 1000. | -37. | 1369. | |
| 5 | 1048. | 32. | 1024. | |

F1 =sum(e[])

You can also think of $(\text{residual})^2$ as the area of a geometric square whose side has length equal to the absolute value of the residual. These geometric sqares are drawn on the graphs below.

**Linear Model 1**

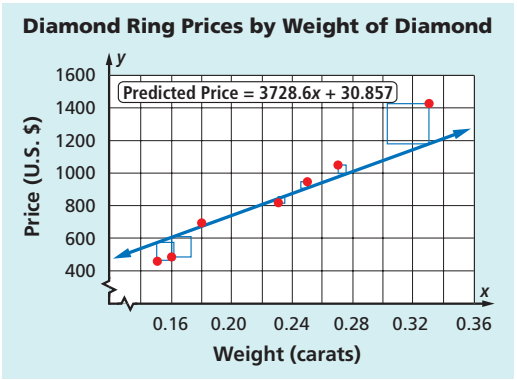
**Squares are shown for a line that does not go through any data points.**




**Total area of the squares ≈ 237,800**

**Linear Model 2**

**Squares are shown for a line through two of the data points.**




**Total area of the squares ≈ 59,870**

The second line is a better model of the data because it has a smaller total area of the squares. The total area is the **sum of squared residuals**.

**Definition of Sum of Squared Residuals**

**Sum of squared residuals** $= \sum_{i=1}^{n} (\text{observed } y_i - \text{predicted } y_i)^2$

The sum of squared residuals is a statistic that measures lack of fit. If you compare two lines, the one with the larger sum of squared residuals is not as good a model as the one with the smaller sum of squared residuals. If you have many possible lines you could use to model data, compute the sum of squared residuals for each model. The line that gives the smallest value provides the best fit to the data.

## Activity

**This table shows the number of televisions per 100 people in 1997 and the number of unemployed per 100 people in 2008 for nine countries from around the world. Some people suggest that increased TV viewing leads to a less productive workforce. Are these statistics related?**

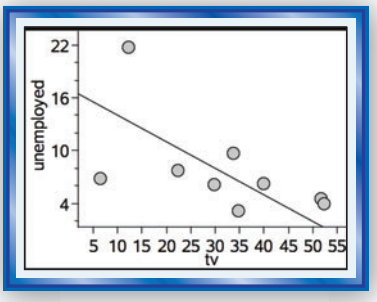
| Country | TVs per 100 | Unemployed per 100 |
|---|---|---|
| Argentina | 22.3 | 7.8 |
| Bulgaria | 40.0 | 6.3 |
| India | 6.5 | 6.8 |
| Israel | 29.9 | 6.1 |
| Netherlands | 51.8 | 4.5 |
| New Zealand | 52.3 | 4.0 |
| Poland | 33.7 | 9.7 |
| South Africa | 12.3 | 21.7 |
| South Korea | 34.7 | 3.2 |

Source: UNESCO Institute for Statistics, CIA World Fact Book

**Step 1** Use a statistics utility to create a scatterplot with number of TVs on the $x$-axis and number of unemployed people on the $y$-axis.

**Step 2** Add a movable line to the scatterplot.

**Step 3** Move the line as close to $y = -0.3x + 17$ as you can. Your screen should look similar to the one at the right. Record the sum of squared residuals.




**Step 4** Move the line so that it goes through the points for Bulgaria and the Netherlands. Record the equation. Record the sum of squared residuals.

**Step 5** Tell which line is the better model of the data. Explain your choice.

**Step 6** Move the line until you get the smallest sum of squares that you can. Record the equation and the sum of squared residuals.

## Questions

### COVERING THE IDEAS

1. Define *linear function*.
2. Refer to the diamond price data in Example 1.
   a. Find an equation for the line that passes through the data points (0.35, 1629) and (0.12, 334.50).
   b. Write a sentence that states how cost increases for every carat increase in weight according to your equation.
   c. Use your line to predict the price of a 0.17-carat diamond.
   d. Is this interpolation or extrapolation?
3. **Fill in the Blanks** For a data set, residuals are the differences between ___?___ values and ___?___ values.

4. When a residual is positive, is the observed value higher or lower than the predicted value?

**In 5 and 6, a diamond speculator used the line with equation $y = 5000x - 250$ to estimate the price of diamond rings.**

5. a. What would the speculator predict for the price of the 0.29-carat diamond ring?
   b. According to the data in Example 1, what is the residual for the 0.29-carat diamond ring?
6. a. What would the speculator predict as a price for the 0.21-carat diamond ring?
   b. What is the residual for this diamond ring?
7. Does the phrase "sum of squared residuals" mean "first sum the residuals and then square the sum" or "first square each residual and then sum the squares?"
8. Consider the linear model $f(x) = 5.2x - 3$.
   a. Given a residual of –0.2 at $x = 1$, find the observed value.
   b. Given a residual of 2.1 at $x = 4$, find the observed value.
   c. Given a residual of 0 at $x = 3$, find the observed value.

## APPLYING THE MATHEMATICS

**In 9 and 10, suppose Jane asked members of her family how many states they had visited. Her data are in the table at the right.**

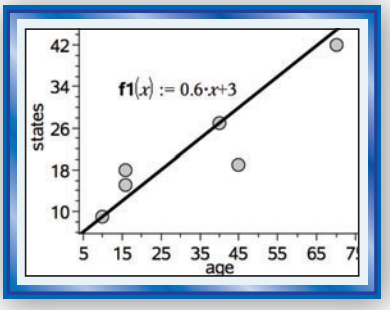
| Family Member | Age | States Visited |
|---|---|---|
| Jane | 16 | 15 |
| Cousin Xia | 16 | 18 |
| Dad | 40 | 27 |
| Grandma | 70 | 42 |
| Brother Ed | 10 | 9 |
| Uncle Ralph | 45 | 19 |

9. Jane plotted the data and drew a line through the points for Ed and Grandma. She found its equation to be $y = 0.55x + 3.5$.
   a. What is the slope of this line and what does it represent?
   b. Complete the spreadsheet below to calculate the residuals for this model.
   c. What is the sum of squared residuals for Jane's model?
   d. Jane used her equation to predict how many states she will have visited at age 30. Is this interpolation or extrapolation?
   e. What is Jane's prediction from Part d?

| ◇ | A | B | C | D |
|---|---|---|---|---|
| 1 | Age | States Visited | Predicted | Residual$^2$ |
| 2 | 16 | 15 | 12.3 | 7.29 |
| 3 | 16 | 18 | | |
| 4 | 40 | 27 | 25.5 | 2.25 |
| 5 | 70 | 42 | 42.0 | 0 |
| 6 | 10 | 9 | | |
| 7 | 45 | 19 | 28.25 | 85.56 |

10. Xia thought that drawing the line through the points for Ed and Dad would look better because then there would be two points above the line and two below the line. Her line has equation $y = 0.6x + 3$.

a. What is the residual for Grandma using Xia's model?
b. What is the squared residual for Dad using Xia's model?
c. Calculate the sum of squared residuals for Xia's model.
d. Xia uses the equation to predict how many states Grandma will have visited when she reaches 100. Is this interpolation or extrapolation?
e. What is Xia's prediction from Part d?




**11.** The table at the right gives expected life spans and gestation periods for selected animals. *Gestation* for mammals is development time in the mother's uterus until birth. For humans, the mean time is 266 days.

| Animal | Expected Life Span (years) | Mean Gestation (days) |
|---|---|---|
| Black Bear | 20 | 215 |
| Camel | 40 | 390 |
| Elephant | 55 | 660 |
| Gorilla | 50 | 251 |
| Gray Squirrel | 9 | 42 |
| Hippopotamus | 45 | 240 |
| Lion | 12 | 100 |
| Moose | 27 | 240 |
| Wolf | 10 | 63 |

Source: BBC

a. Make a scatterplot with Life Span as the independent variable.
b. Draw the line through the points for Lion and Elephant. Do you think this line fits the data well?
c. Find an equation for the line in Part b in slope-intercept form.
d. Interpret the slope of the line in Part b.
e. What is the residual for the Moose under this model?
f. Use the line in Part b to predict the gestation period for the Blue Whale, which has an estimated life span of 80 years. (The actual gestation period is estimated at 11-12 months.) Is this interpolation or extrapolation?
g. What is the sum of squared residuals for the line?

## REVIEW

**12.** Consider the function {(–2, 8), (–1, 8), (0, 8), (3, 8), (8, 8)}. **(Lesson 2-1)**
a. State its domain.
b. Give its range.

**In 13 and 14, a sample of five math test scores is given. (Lesson 1-6)**
a. **Find the standard deviation of the data set.**
b. **Explain a conclusion you can draw about the spread of the data set by just looking at it.**

**13.** 63, 77, 81, 83, 92

**14.** 94, 95, 95, 96, 100

## EXPLORATION

**15.** Use statistical software to draw a line that has the smallest sum of squared residuals for the data in the table in Question 11.
a. What is the equation?
b. What is the sum of squared residuals?

**QY ANSWER**

$225.50