Lesson

2-8 Choosing a Good Model

Vocabulary

residual plot

BIG IDEA The residual plot can help you choose a model for a given data set.

You have seen how to fit linear, exponential, quadratic, and inverse variation models to data. One aspect of modeling, however, is worthy of further discussion: *How do you know you have found a good model?* If two models appear to fit the data equally well, it is usually wise to pick the simpler of the two, but how can you tell how well data fit a model? One measure of how well a model fits the data is the correlation coefficient, but this applies only to linear models.

Another method to determine how well a model fits data is to analyze the residuals. When you evaluate a model at the x-value for a particular data point, you are likely to get a predicted y-value that is different from the observed y-value. Recall that the residual is the difference.

residual = error = observed y-value − predicted y-value

Mental Math

Name the figure that is the graph of the equation.

a. $3xy = 4$

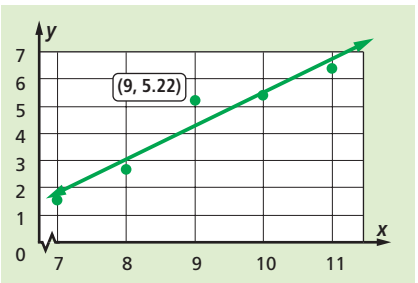
b. $3x + y = 4$

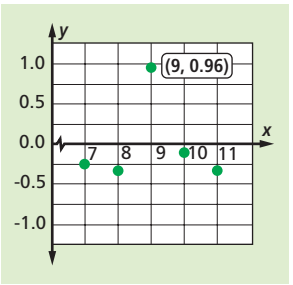
c. $3x^2 + y = 4$

d. $y = 3 \cdot 4^x$

Residual Plots

The graph at the left below shows a scatterplot and linear regression model for a data set. The graph at the right below shows the *residual plot* for that model. The point (9, 5.22) marked on the scatterplot has its corresponding point marked on the residual plot, (9, 0.96). This means that the data point $x = 9$ has an observed value of 5.22 and a residual value of 0.96 under the linear regression model.



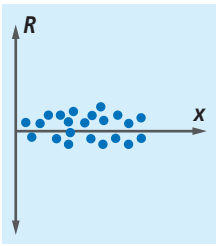
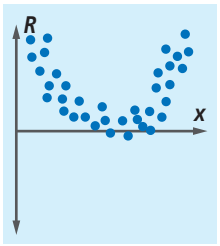
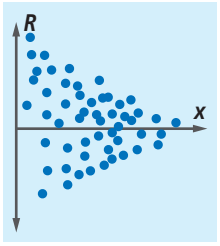


STOP QY

QY

What is the predicted value for $x = 9$?

A **residual plot** pairs each x-value from the data set with its residual. If the residuals are clustered around the x-axis, as shown in the leftmost diagram below, the model is likely to be a good fit for the data. If, however, the residuals have a different pattern, then a better model can probably be found. The second and third graphs below show patterns of residuals indicating that both models need improvement.

Residual Pattern:	clustered around x-axis	same side of axis at ends	shaped like funnel
Residual Plot:			
	good model	weak model	weak model

Analyzing Residuals for a Linear Model

Activity 1

The table at the right shows the length L of each day (sunrise to sunset) observed at a city in the Northern Hemisphere every 10 days for 100 days beginning on August 31st. D is the number of days after this date.

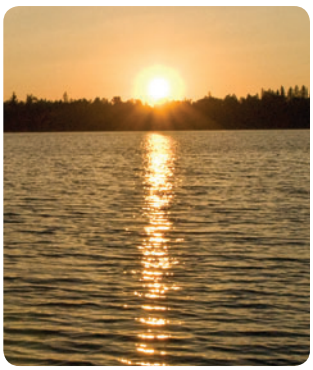
D (days)	L (minutes)
0	793
10	766
20	739
30	711
40	684
50	657
60	631
70	607
80	586
90	568
100	556

Step 1 Use a statistics utility to create the scatterplot of the data.

Step 2 Find the regression line and correlation coefficient for these data. What does the correlation coefficient indicate about a linear model for these data?

Step 3 Reproduce the spreadsheet below. Then use your calculated regression line to fill in both the predicted day lengths and the associated residuals. A few entries have already been filled in.

◇	A	B	C	D
1	Days (D)	Observed Day Length (L)	Predicted Day Length (p)	Residual (R = L – p)
2	0	793	786.05	6.95
3	10	766	761.53	4.47
4	20	739		
5	30	711		-1.49
6	40	684		
7	50	657		
8	60	631		
9	70	607		
10	80	586		
11	90	568		
12	100	556		



Sunset on Lake Franklin in the Chequamegon-Nicolet Forest in Northern Wisconsin

(continued on next page)

Step 4 Plot the residual set of points (D, R) on the same axes as the scatterplot in Step 1.

Step 5 Based on the residual plot, what do you conclude about a linear model for these data?

Activity 1 tested the theory that the number of minutes of daylight decreases in a linear fashion. Even though the correlation coefficient in Step 2 indicates that a line is a good model, the residuals show that there is a better model. Therefore, the researcher must seek another theory or more realistic model to explain the manner in which daylight decreases.

One way to seek a better model is to gather more data. In Activity 1, the hours of sunlight were provided for about 3 months of the year. This is not the full domain of the situation. Data for two or more years would show a periodic wavy pattern that requires functions you will study in later chapters of these book.

Analyzing Residuals for Other Models
Exponential Model

Activity 2

Step 1 Use a statistics utility to create a scatterplot of the U.S. Census data shown at the right. Use years after 1790 as the independent variable.

Step 2 Compute the exponential regression model for these data.

Step 3 Using the exponential regression model, compute the predicted populations and residuals. Organize the data in a spreadsheet like the one shown below.

Step 4 Plot the set of residual points (Y, R).

Step 5 Based on the residual plot, why is the exponential model not a good fit for these data?

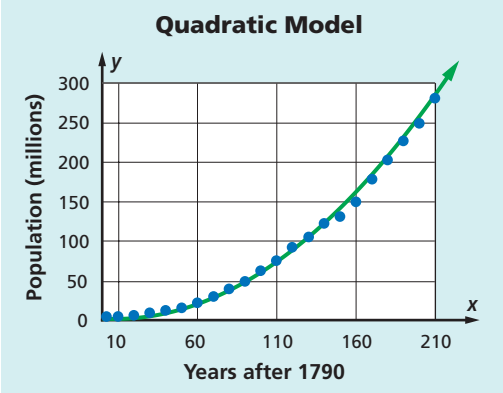
Year	Population (millions)	Year	Population (millions)
1790	4	1900	76
1800	5	1910	92
1810	7	1920	106
1820	10	1930	123
1830	13	1940	132
1840	17	1950	151
1850	23	1960	179
1860	31	1970	203
1870	40	1980	227
1880	50	1990	249
1890	63	2000	281

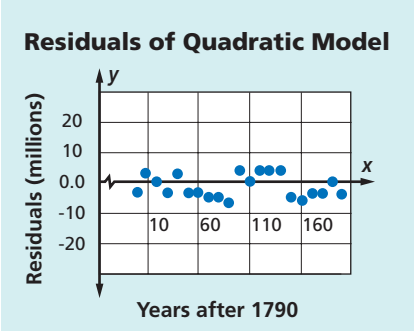
Source: U.S. Census Bureau

◇	A	B	C	D	E
1	Year	Year – 1790 = (Y)	Population (m)	Predicted Population (p)	Residual (R = m – p)
2	1790	0	4	6.03	-2.03
3	1800	10	5	7.39	-2.39
4	1810	20	7	9.04	-2.04

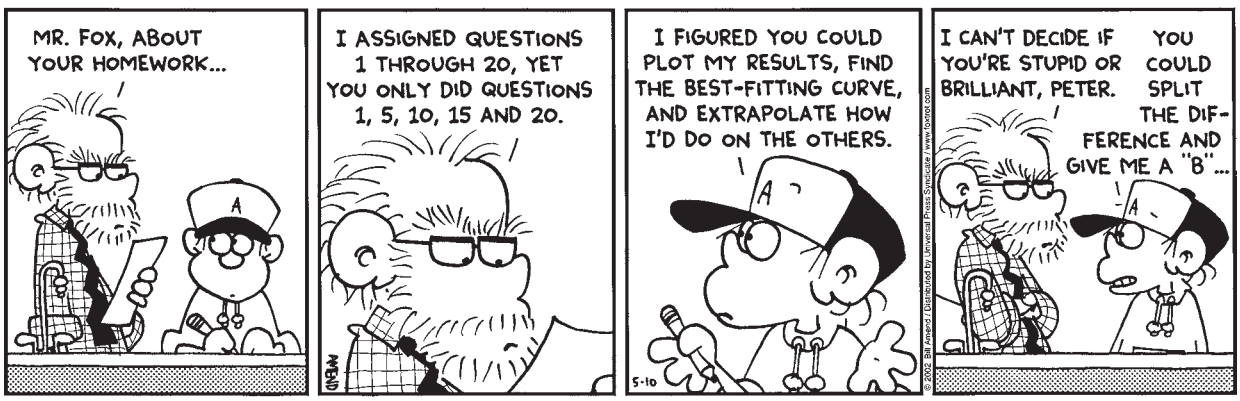
Activity 2 suggests that although exponential regression models are useful for modeling population growth, they frequently break down over time. Limited resources and other factors prohibit populations from growing indefinitely.

A graph of the quadratic regression model of the U.S. population data, $y = 0.007x^2 - 0.1138x + 6.1097$, is shown at the left below. It turns out that this quadratic model is a much better fit than the exponential model. The residuals are relatively small and they cluster in a horizontal band centered around the x-axis. However, the quadratic model is an impressionistic model of population growth because there is no theory that supports a quadratic relationship between year and population.





As mentioned in earlier lessons, extrapolation is risky business. This is particularly true when there is no theory to support the model. The quadratic model is a very good model for the population of the U.S. from 1790 to 2000, but there is no assurance that the model will make accurate predictions for years outside the data set. An even better model than the quadratic model might be a piecewise function, with each piece chosen to best fit a portion of the domain.



Careful Modeling

You should consider at least five things when building a model from data. Assuming that your data are a representative sample of the population of interest, you should:

1. Build a model from theory, if possible. Some real-world situations suggest certain models.
2. Graph the data on a scatterplot. Draw the model on the same graph. A good model should follow any pattern or trend in the data.

3. Graph the residuals. If the residual plot does not fall within a relatively narrow horizontal band centered around the x-axis or if there is a pattern to the residuals, you may have to change your theory or look for a better model.
4. Use the correlation coefficient to check whether a linear model is appropriate.
5. In all cases, be aware of the model's ability or limitation for interpolation and extrapolation.

Sometimes There Is No Good Model

Everyone who invests in the stock market wants to buy stocks when their prices are low and sell when their prices are high. The difficulty is that when you buy today because you think the price is low, you have no guarantee that the price tomorrow will be higher. In many cases the price goes down and your investment loses value. In the same way, you might sell a stock today because you think prices are high and likely to go down, only to find out that the prices are even higher tomorrow. No model has been developed that can accurately predict changes in the stock market, but many people make a living by developing models that work for a short time or in special situations.

The graph below shows the Dow Jones industrial average from July, 2001 to March, 2009.

Source: Yahoo! Finance 2008

The graph shows many individual trends, but no overall, consistent trend. There are many fluctuations. Catastrophic events often cause significant changes in the graph.

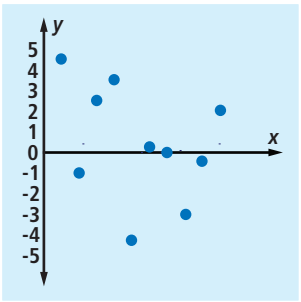
One example of this was the destruction of the World Trade Center on September 11, 2001, which initiated a drop in the stock market. (See if you can locate the drop in the Dow-Jones average due to September 11 on the graph.) Even the most sophisticated stock traders with powerful statistical models were caught off guard by this event.

Questions

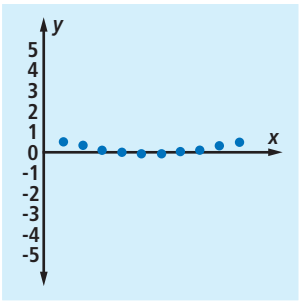
COVERING THE IDEAS

1. What characterizes a residual plot of a good model?
2. **Multiple Choice** Look at the residual plots below. Which indicates the best model? Explain your answer.

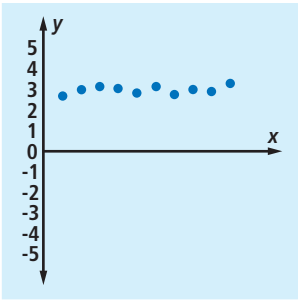
A



B



C



3. a. An exponential regression model for the length L of the Dth day in Activity 1 is $L = 792.63(0.996312)^D$. Compute a table of predicted values using this model.
   b. Compute a table of residual values using the results from Part a.
   c. Create a plot of the residuals from Part b and explain if the exponential regression model is a good model for this data.

In 4–6, refer to Activity 2.

4. Why is it reasonable to expect that an exponential model is an appropriate theoretical model for the population data?
5. a. Find an exponential regression model for the population data using only the years from 1790 to 1880.
   b. Make a spreadsheet with year, census population, predicted population, and residual.
   c. Plot the residuals.
   d. Is an exponential model appropriate for the 1790–1880 time period? Explain your decision.
6. a. Use the quadratic regression equation to predict the 1900 population to the nearest million.
   b. Use the value in Part a to calculate the residual for the 1900 population.
7. Why do you think no good model has been found for predicting the future prices of stocks?

APPLYING THE MATHEMATICS

8. The chambered nautilus is a cephalopod mollusk, a relative of octopuses and squids. It creates a spiral shell that has inspired mathematicians and poets for centuries. As it grows, the animal partitions off increasingly larger cells to inhabit. The table below gives the volume (in cubic centimeters) of chambers 10 through 20 of a nautilus.

Chamber number	10	11	12	13	14	15	16	17	18	19	20
Volume (cm^3)	0.2	0.5	0.6	0.6	0.8	0.9	1	1.5	1.7	2.1	2.5

Source: Paleobiology

   a. Is a linear or exponential function the more appropriate theoretical model for the data? Give a reason for your answer.
   b. Make a scatterplot of the data, using chamber number as the independent variable. From the scatterplot, which model seems more appropriate?
   c. Determine whether the linear or exponential regression model better fits the data using residual plots as well as the value of the linear model's correlation coefficient.
   d. Write a sentence indicating which model you would choose, and why.

9. The table at the right shows how many millions of miles motor vehicles drove in the U.S. in certain years.

Year	Miles (millions)
1920	47,600
1930	206,320
1940	302,188
1950	458,246
1960	718,762
1970	1,109,724
1980	1,527,295
1990	2,144,362
2000	2,746,925

Source: Historical Statistical Abstract

   a. Is any theoretical model appropriate? If so, what kind and why?
   b. Make a scatterplot of the data. What kind of model seems appropriate?
   c. Sketch the residual plots for the linear, exponential, and quadratic regression models.
   d. Write a sentence indicating which model you would choose and why.

10. Seasonally-adjusted U.S. unemployment figures for June of each year are given in the table below.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Unemployed (thousands)	6799	6212	5951	5651	6484	8393	9266	8280	7536	7017	6997	8499

Source: Bureau of Labor Statistics

   a. Make a scatterplot of the data, plotting the number of years after 1997 as the independent variable.
   b. Which, if any, of a linear, exponential, or quadratic function seems to model these data? Justify your answer.

REVIEW

11. Recall that the force between electrical fields is inversely proportional to the square of the distance between them. When two electrically charged particles are 0.8 m apart, the force between them is 5 N. What will the force between them be when they are 0.1 m apart? **(Lesson 2-7)**

12. **Multiple Choice** If grades g in school vary inversely as the number t of hours spent watching TV every day, then which of the following describes the relationship between g and t? **(Lesson 2-7)**

A $g = kt^2$ **B** $g = \frac{k}{t}$

C $g^2 = \frac{k}{t}$ **D** $g = \frac{k}{t^2}$

13. The population of Florida from 1950 to 2000 is given in the table below. **(Lesson 2-6)**

Year	1950	1960	1970	1980	1990	2000
Population (thousands)	2,771	4,952	6,791	9,747	12,938	15,982

Source: Historical Statistical Abstract

a. Make a scatterplot of the data, plotting number of years after 1950 as the independent variable.

b. Find a quadratic regression model for the data.

c. Use your model to predict the population of Florida in 2015.

Tampa, Florida

14. A 6" diameter frozen personal cheese pizza lists its total calories as 590. About how many calories are in a single 2-square-inch bite of pizza? **(Previous Course)**

EXPLORATION

15. Find population data over a number of years for your city or state. Find the linear, quadratic, or exponential model that best fits the data. Use your model to predict the population of your city or state in the year 2050. Do you believe this a reasonable prediction? Explain your response.

QY ANSWER

4.26