A Fitting Explanation

How can you tell when forecasters have made reasonable predictions for the future? In this module, you use regression to fit curves to data sets and evaluate models.

Terri Dahl • David Thiel • Deanna Turley

A Fitting Explanation

Introduction

In previous modules, you've used various types of equations to model data, understand patterns, and make predictions. To evaluate models, you've used graphs and residuals. In this module, you explore some tools for evaluating linear models.

Activity 1

Before you can analyze how well a linear model describes a data set, you must first suggest the model. In this activity, you create some simple models, then begin to analyze them.

Exploration 1

In this exploration, you use equations to model the relationship between the age and length of fish. Table **1** gives the age and length of a random sample of 20 fish from a population of trout.

Table 1: Ages and lengths of 20 trout

Age (yr)	Length (mm)	Age (yr)	Length (mm)
1	65	3	295
1	72	3	356
1	93	4	355
1	103	4	487
2	209	4	443
2	173	4	371
2	148	5	507
2	181	5	423
3	324	5	398
3	401	5	551

a. Use a graphing utility and the data in Table **1** to create a scatterplot of length versus age.

Mathematics Note

The mean is one of the simplest ways to describe a set of one-variable data. Similarly, the mean line is one of the simplest models for a set of two-variable data.

The equation of the **mean line** for a set of data points in the form (x,y) is $y = \bar{y}$, where $\bar{y}$ represents the mean of the y-values.

For example, consider a data set consisting of the points (1,2), (2,4.1), and (3,5). The mean of the y-values ($\bar{y}$) is 3.7. Therefore, the equation of the mean line for this data set is $y = 3.7$.

b. While the mean line may not always provide an appropriate model for a data set, it can be useful for comparing other potential models.

1. Determine the equation of the mean line for the data in Table **1**.

2. Graph the mean line on the scatterplot from Part **a**.

c. Another possible model for a data set is a linear regression equation.

1. Use technology to determine the linear regression equation for the data in Table **1**. To distinguish this model from the mean line, designate this equation y' and record it in the form $y' = mx + b$.

2. Graph the linear regression equation on the scatterplot from Part **a**.

d. Predict the length of a 5-year-old trout using each of the following models:

1. the mean line

2. the linear regression equation.

Discussion 1

a. Does the data in Table **1** appear to show any kind of association between the age of the fish and the length of the fish? Explain your response.

b. Which appears to be a better model for this data: the mean line or the linear regression equation? Explain your response.

c. If you caught a 5-year-old trout, would you expect its length to be exactly as predicted by either model? Explain your response.

Exploration 2

Figure **1** shows a scatterplot of the data from Table **1**, the mean line for the data, and a graph of the corresponding linear regression equation. Notice that the linear regression equation provides some information about the association between the two variables in the data, while the mean line describes only a central tendency.

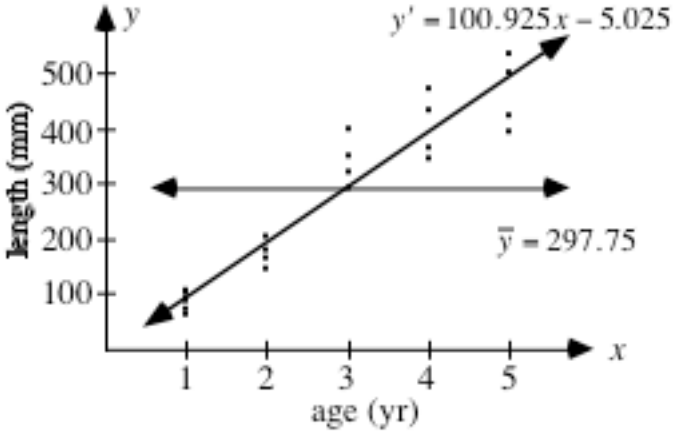


Figure 1: Scatterplot, mean line, and regression equation

a. Figure **2** below shows graphs of the mean line and linear regression equation, along with the data point (5,551). In this case, the deviation of the predicted value $y' = 499.6$ from the mean $\bar{y} = 297.75$ can be attributed to the positive association in the data.

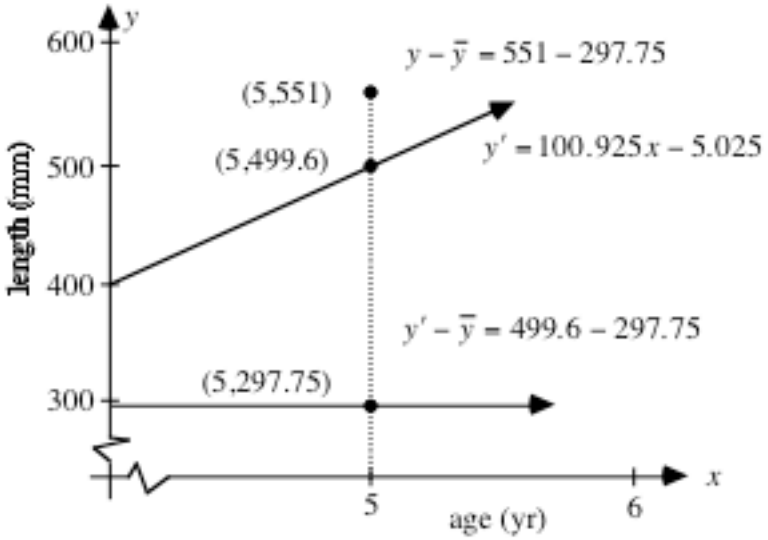


Figure 2: Data point (5,551) with mean line and regression line

1. In Figure **2**, determine the deviation from the mean line of the value predicted by the linear regression model, or $y' - \bar{y}$.
2. Determine the deviation from the mean line of the actual *y*-value of the data point, or $y - \bar{y}$.

3. If a regression model exactly fits a set of data, then there is no deviation of a data point from the regression line. In other words, $y' = y$. Using the mean line as a reference, this can be expressed as:

$$\frac{y' - \bar{y}}{y - \bar{y}} = 1.00$$

This indicates that 100% of the deviation from the mean line is explained by the regression equation.

For the data point (5,551), what percentage of the deviation from the mean line is explained by the linear regression model?

Mathematics Note

The **total variation** for a data set is the sum of the squares of the deviations from the mean line of the data points.

The **explained variation** for a data set is the sum of the squares of the deviations from the mean line of the values predicted by the linear regression equation.

For example, Table **2** shows the distance traveled by a truck for several different amounts of gasoline. In this case, $\bar{y} = 50.65$ and the linear regression model is approximately $y' = 8.33x - 0.09$. As shown in the table, the total variation is 258.37, while the explained variation is 257.78.

Table 2: Total variation and explained variation for a data set

Liters of Gas (x)	**Kilometers Traveled (y)**	**Linear Regression Model (y')**	**Explained Variation $(y' - \bar{y})^2$**	**Total Variation $(y - \bar{y})^2$**
2.50	20.75	20.74	894.66	894.01
6.25	52.00	51.98	1.78	1.82
6.25	52.50	51.98	1.78	3.42
7.35	59.75	61.15	110.21	82.81
8.10	68.25	67.40	280.45	309.76
		Sum	257.78	258.37

The **coefficient of determination (r^2)** is the percentage of the total variation from the mean line explained by the linear regression equation, or:

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

The value of r^2 is often reported as a decimal. It represents the proportion of the total variation in the *y*-values that can be explained by the linear relationship in the data set.

For example, the coefficient of determination for the data in Table **2** and the linear regression $y' = 8.33x - 0.09$ is:

$$r^2 = \frac{257.78}{258.37} \approx 0.9989$$

This means that approximately 99.89% of the variation from the mean line is explained by the linear relationship between the liters of gas used and the kilometers traveled by the truck.

b. Calculate the total and explained variation for the data in Table **1**.

c. Calculate r^2, the coefficient of determination, for the data in Table **1**.

Discussion 2

a. Describe what the coefficient of determination calculated in Part **c** of Exploration **2** represents in terms of the ages and lengths of the trout in Table **1**.

b. Why do you think the deviations are squared when determining total variation and explained variation?

Mathematics Note

The statistical association between two variables is referred to as **correlation**.

The **linear correlation coefficient (*r*)** is found by taking the square root of the coefficient of determination r^2. The value of r is in the interval $[-1,1]$, where $|r| = \sqrt{r^2}$. The slope of the linear regression model determines if r is positive, negative, or zero.

The linear correlation coefficient is a measure of how closely the points in a data set can be modeled by a line. The closer $|r|$ is to 1, the stronger the linear relationship between the variables. When $r = 0$, the *y*-values in a data set are said to have no linear correlation to the *x*-values.

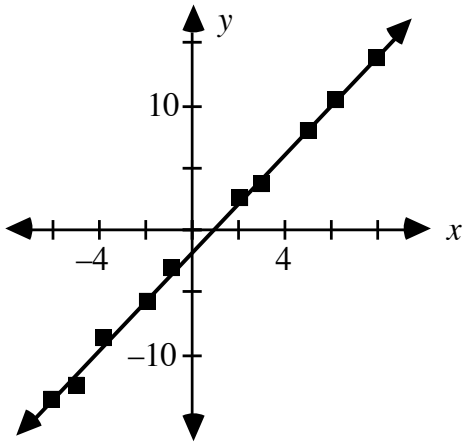
For example, consider the data points (1,2.7), (4,1.2), (7,–0.3), (10,–0.8), and (13,–2.3). The regression equation for this data set is $y = -0.4x + 2.9$. Since the slope of the regression equation is negative, the value of r is negative. Since $r^2 \approx 0.9850$, then $r \approx -0.99$. This indicates a strong negative linear relationship between the variables *x* and *y*.

c. **1.** What does it mean when *r* indicates a strong negative relationship between *x* and *y*?

2. What does it mean when *r* indicates a strong positive relationship between *x* and *y*?

Assignment

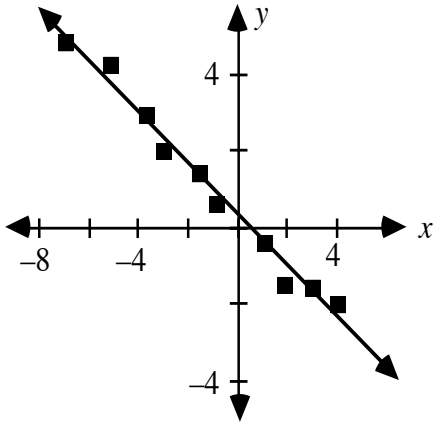
1.1 **a.** The graph below shows the linear regression equation for a data set.



1. Explain how the sign of the linear correlation coefficient r can be determined by examining the graph.

2. Given that the coefficient of determination $r^2 \approx 0.996687$, determine r.

b. The following graph shows the linear regression equation for another data set.



1. Explain how the sign of the linear correlation coefficient r can be determined by examining the graph.

2. Given that the coefficient of determination is $r^2 \approx 0.980847$, determine r.

c. Describe how to determine the linear coefficient r, in general, given the coefficient of determination r^2.

1.2 The following table shows the distance traveled by a hiker for several different periods of time.

Time (hr)	Distance (km)
2.50	12.00
6.25	30.00
7.35	35.28
8.10	38.88

a. Determine the linear regression equation that models a scatterplot of distance versus time.

b. Calculate the coefficient of determination, r^2, and describe what this value indicates in terms of the times and distances in the table.

c. Is there a strong linear relationship between time and distance? Explain your response.

1.3 The following table lists the prices of telescopes of comparable quality, along with their lens diameters.

Lens Diameter (in inches)	Price of Telescope
6	$295.00
8	$395.00
10	$565.00
12.5	$765.00
16	$995.00

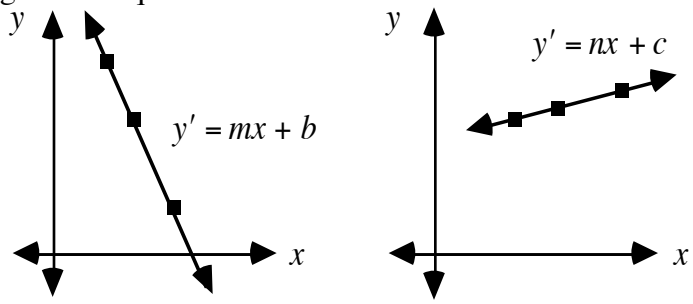
a. Determine the linear regression equation that models a scatterplot of price versus lens diameter.

b. Use your model to predict the price of a 9-inch telescope.

c. Find the linear correlation coefficient, r, for the data set.

d. Describe what the linear correlation coefficient means in terms of the lens diameters and prices given in the table.

1.4 The following graphs show two sets of data modeled by linear regression equations:



a. Assuming that the scales on the graphs are identical, compare the total variations of the two data sets.

b. Given that, in both cases, 100% of the total variation is explained by the linear relationship between the *x*- and *y*-values in the data set, determine the corresponding linear correlation coefficients.

1.5 The table below shows some information on the price of telescope mirrors of comparable quality with respect to their diameters.

Diameter (in inches)	Price of Mirror
1.30	$26.95
1.52	$35.95
1.83	$37.95
2.14	$45.95
2.60	$84.95

a. Find the equation of the regression line for this data set.

b. Calculate the coefficient of determination, r^2.

c. What does your response to Part **b** indicate in terms of the prices and diameters of telescope mirrors given in the table?

d. Is there a strong linear relationship between the diameter of a telescope mirror and its price? Explain your response.

* * * * *

1.6 Drew mows lawns in the summer. As part of his weekly accounts, he records both the number of hours worked and the liters of gasoline used. This data is shown in the table below.

Week	Time (hr)	Gasoline (L)
1	32	14.2
2	31	13.3
3	25	12.0
4	35	15.0
5	46	20.2
6	28	12.6
7	34	14.8

a. Determine a linear regression model for a scatterplot of gasoline used versus hours worked.

b. Determine r^2, the coefficient of determination.

c. Describe what r^2 indicates in terms of the hours spent mowing and the amount of gasoline required.

d. Use the model from Part **a** to estimate the amount of gasoline required to mow lawns for 36 hr.

e. Would you expect your estimate from Part **d** to be accurate? Explain your response.

* * * * * * * * * *

Activity 2

A linear correlation coefficient r near either –1 or 1 implies a strong linear relationship. However, when $|r|$ for a linear regression is close to 0, it does not necessarily indicate that there is *no* relationship at all between the variables in the data. It means only that there is no linear relationship.

When analyzing data with regression lines, it is also important to avoid the conclusion that a strong correlation implies a cause-and-effect relationship. For example, the correlation coefficient for a set of data showing average annual salary versus number of years in school is close to 1. This indicates a positive linear relationship between these two quantities. As the number of years in school increases, so does the average salary. However, simply attending school does not *cause* someone to earn a large salary. In this activity, you explore some of the limitations of linear regression models.

Exploration

The data in Table **3** below shows the braking distance required to stop a car moving at various speeds. **Note:** Braking distance varies according to the mass of the car, the condition of the brakes, the road conditions, and other factors. The data below is for one type of car with new brakes on a dry road.

Table 3: Braking distance for various speeds

Speed (km/hr)	Braking Distance (m)
10	0.475
15	1.069
20	1.900
25	2.969
30	4.275
35	5.819

a. Use the data in Table **3** to create a scatterplot of braking distance versus speed and find the equation of the regression line.

b. Determine the coefficient of determination, r^2.

c. Use your graph of the data set, the regression equation, and r^2 to assess how well the equation models the trend in the data.

Mathematics Note

A **residual** is the difference between the *y*-value of a data point and the corresponding *y*-value of the model. In other words, it is the difference between an observed value and the predicted value.

The **average deviation of prediction** is the square root of the average of the squared residuals. It can be found using the following formula, where *n* is the number of data points:

$$\text{average deviation of prediction} = \sqrt{\frac{\text{sum of the squares of the residuals}}{n}}$$

For many data sets, the average deviation of prediction can be used to measure the reliability of a prediction. As a rule of thumb, it is likely that the *y*-values of most data points with *x*-values near the mean ($\bar{x}$) fall within 2 average deviations of the regression line. Thus, it is likely that the predicted *y*-value will be within 2 average deviations of the true *y*-value. The interval of 2 average deviations of prediction on either side of the predicted *y*-value is an **approximation interval**.

For example, the data for trout lengths and ages in Table **1** can be modeled by the regression line $y' = 100.925x - 5.025$. Using this model, the sum of the squares of the residuals is 51,012. Since there are 20 data points, the average deviation of prediction is:

$$\sqrt{51{,}012/20} \approx 50.5$$

The mean of the *x*-values in Table **1** is 3 yr. Using the regression line, the predicted length of a 3-yr-old trout is approximately 298 mm. Thus, it is likely that the actual *y*-value associated with $x = 3$ falls in the interval $[298 - 2(50.5), 298 + 2(50.5)]$ or $[197,399]$. In other words, it is likely that the length of a 3-yr-old trout is greater than or equal to 197 mm and less than or equal to 399 mm.

d. Use your model for the data in Table **3** to predict a braking distance, along with the corresponding approximation interval, for each of the following speeds:

1. 22 km/hr

2. 34 km/hr

3. 50 km/hr

4. 75 km/hr.

Discussion

a. How well do you think your model from Part **a** of the exploration describes the trend in the data? Explain your response.

b. Describe how the average deviation of prediction resembles the standard deviation of the *y*-values in a data set.

c. Given the similarities between average deviation of prediction and standard deviation, why would you expect predicted *y*-values to fall within 2 average deviations of the actual *y*-values?

d. The actual braking distances for the speeds given in Part **d** of the exploration are shown in Table **4** below.

Table 4: Braking distances for four different speeds

Speed (km/hr)	Braking Distance (m)
22	2.299
34	5.491
50	11.875
75	26.719

For which speeds did the regression line provide a prediction whose approximation interval contained the actual braking distance? Which of these speeds are within the range of the *x*-values of the original data set in Table **3**?

e. What conclusions can you draw about the reliability of predictions made for values outside the range of a data set?

Assignment

2.1 The following table shows information on motor vehicle registrations and fatalities for eight states.

State	1990 Motor Vehicle Registrations (in thousands)	1990 Motor Vehicle Fatalities
Wyoming	528	125
North Dakota	630	112
South Dakota	650	153
Montana	783	212
Idaho	1054	243
Utah	1206	270
Oregon	2445	578
Washington	4257	825

Source: U.S. Bureau of the Census, 1993.

a. Use this data to create a scatterplot of fatalities versus registrations. Graph the corresponding regression line on the same set of axes.

b. Find the coefficient of determination and describe what it represents in terms of the data.

c. In 1990, Texas had 12,800,000 motor vehicle registrations. Use the regression line to predict this state's number of motor vehicle fatalities.

d. Calculate the average deviation of prediction for your model and determine an approximation interval for your prediction in Part **c**.

e. In 1990, the actual number of motor vehicle fatalities in Texas was 3243. Compare your prediction with this value and suggest some possible reasons for any difference you observe.

2.2 The following table shows the average monthly income in 1993 for various years of education. **Note:** In this case, 12 years of education is equivalent to earning a high school diploma.

Years of Education	Average Monthly Income ($)
12	1380
13	1579
14	1985
16	2625
18	3411
20	4328

Source: U.S. Bureau of the Census, 1995.

a. Use this data to create a scatterplot of average income versus years of education. Graph its corresponding regression line on the same set of axes.

b. Find the coefficient of determination and explain what it represents in terms of the data.

c. Use the regression line to estimate the average monthly income of people with 15 years of education.

d. Determine an approximation interval for your estimate and explain what it represents in terms of the data.

2.3 The following table shows the mean January temperature and mean annual snowfall for 14 U.S. cities.

City	Mean January Temperature (˚C)	Mean Annual Snowfall (cm)
Minneapolis, MN	–11.2	126.5
Mobile, AL	9.9	1.0
Atlantic City, NJ	–0.6	41.9
Omaha, NE	–6.1	76.7
Providence, RI	–2.3	90.7
Raleigh, NC	3.8	18.3
Reno, NV	0.5	62.5
Albuquerque, NM	1.2	27.9
Sacramento, CA	7.3	0.0
Houston, TX	10.2	1.0
Sioux Falls, SD	–10.1	100.1
Spokane, WA	–2.7	127.8
Chicago, IL	–6.1	97.8
Cleveland, OH	–4.1	138.2

Source: U.S. Bureau of the Census, 1993.

a. Create a scatterplot of mean annual snowfall versus mean January temperature.

b. Find the equation of the regression line for the data.

c. Baltimore, Maryland, has a mean January temperature of 0.4 ˚C. Caribou, Maine, has a mean January temperature of –11.8 ˚C. Use approximation intervals to predict a range for the average annual snowfall in each city.

d. Baltimore's average annual snowfall is 55.4 cm, while Caribou's is 287.8 cm. How do these values compare with your predictions?

e. How much of the variation in snowfall appears to be explained by January temperature? Explain your response.

f. Describe some other factors that may influence average annual snowfall other than January temperatures.

2.4 **a.** The following table shows the prices of telescopes of comparable quality and different lens diameters. Graph the data and the corresponding regression line.

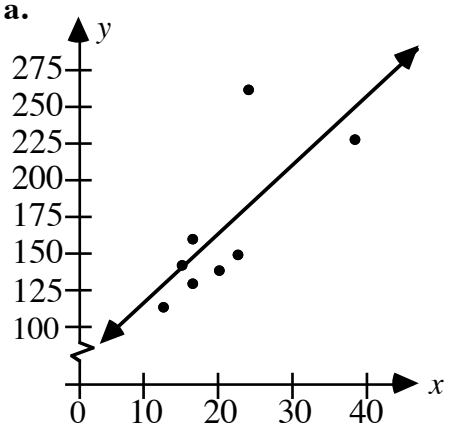
Lens Diameter (inches)	Telescope Price
6	$295.00
8	$395.00
10	$565.00
12.5	$765.00
16	$995.00

b. Use an approximation interval to estimate the cost of a telescope with an 11-inch lens.

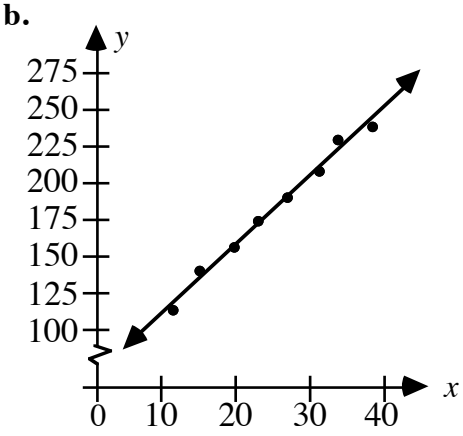
* * * * *

2.5 Each of the two data sets below can be modeled by the same regression equation, $y' = 4.64x - 710$. Compare the approximation intervals for these data sets.

a.



b.



2.6 The magnitude of a celestial body is a measure of its apparent brightness. (As magnitude decreases, apparent brightness increases.) The table below shows the diameter and magnitude of the 10 largest asteroids in the solar system.

Asteroid	Diameter (km)	Magnitude
Ceres	780	7.4
Pallas	489	8.0
Vesta	391	6.5
Hebe	195	8.5
Iris	195	8.4
Juno	190	8.7
Metis	130	8.9
Flora	90	8.9
Astraea	80	9.9
Hygeia	64	9.5

a. Graph a scatterplot of diameter versus magnitude, along with the corresponding regression line.

b. Write a paragraph describing how well the regression equation explains the relationship between diameter and magnitude.

c. **1.** Determine the average deviation of prediction for this model.

2. Calculate an approximation interval for the magnitude of an asteroid whose diameter is 73 km and describe how reliable you think this interval might be.

* * * * * * * * * *

Research Project

Imagine that you are a research statistician. Collect data for two quantities that you believe may have a strong linear correlation. Analyze this data and prepare a presentation of your findings.

Activity 3

As you may recall from Activity **1**, many data sets contain more than one *y*-value for each *x*-value. For example, consider the heights (in centimeters) and ages (to the nearest year) of students in your school. Since there are many different heights for each age, no function can describe every point in this data set.

In the following exploration, you use this type of data to continue your investigation of linear correlation.

Exploration

a. Obtain a stopwatch from your teacher and use it to time several different events. Compare the time you recorded for a specific event with the time recorded by a classmate. How accurate are your measurements?

b. **1.** Devise a method for approximating time intervals of 5 sec, 10 sec, 15 sec, and 20 sec without using a clock or stopwatch.

2. Use a stopwatch to measure how accurately your method determines the times you intended. Record four measurements for each interval in a table with headings like those in Table **5**.

Table 5: Time measurements

5 sec	10 sec	15 sec	20 sec

c. **1.** Create a scatterplot of measured time versus the intended time and determine the equation of the regression line.

2. Determine the linear correlation coefficient.

3. Determine the average deviation of prediction.

d. Find the mean of the values in each column in Table **5**. Compare these means to the values predicted by the regression line.

Discussion

a. Do you think the regression line is a good model for describing the data in the exploration? Explain your response.

b. In situations where a data set has many y-values for each x-value, statisticians often use regression equations to estimate the mean of the y-values for each x-value.

Does the regression line from Part **c** of the exploration appear to be more useful for predicting the y-values for each x-value or for predicting the mean of the y-values? Explain your response.

c. Given an intended time of 50 sec, do you think that the regression line would provide a good prediction for the measured time? Explain your response.

Assignment

3.1 The table below lists the shoe sizes and heights of 10 different people.

Shoe Size	Height (cm)
8	162
8	178
9	178
9	172
10	180
10	178
13	183
13	188
15	196
15	193

a. Find a linear regression equation that could be used to predict a person's height given that person's shoe size.

b. Using your model from Part **a**, determine an approximation interval for the heights of people with a shoe size of 11.

c. Would you expect your approximation interval to contain the heights of all people with the given shoe size? Explain your response.

3.2 In Activity **1**, you determined a linear regression model for the following data on the age and length of trout.

Age (yr)	Length (mm)	Age (yr)	Length (mm)
1	65	3	295
1	72	3	356
1	93	4	355
1	103	4	487
2	209	4	443
2	173	4	371
2	148	5	507
2	181	5	423
3	324	5	398
3	401	5	551

a. Determine the mean length of trout of each age in the data set.

b. Does the linear regression model for the entire data set provide reasonable predictions for the mean length of trout at a given age? Explain your response.

3.3 The table below shows the engine size in liters and fuel economy in miles per gallon for 10 different cars.

Engine Size (L)	Fuel Economy (mpg)
2.5	22
2.5	24
1.9	27
1.9	29
3.3	19
3.3	20
1.6	29
1.6	29
5.7	17
4.3	15

a. Use this data to create a scatterplot of fuel economy versus engine size and describe any association you observe.

b. Determine the linear regression equation for the data and calculate r^2, the coefficient of determination.

c. Use the model to predict the fuel economy of a car with a 2.8-L engine. Include an approximation interval with your prediction.

d. Explain whether or not you believe the model can accurately predict the fuel economy for a given engine size.

* * * * *

3.4 The following table shows the price and reliability ratings of 10 new cars. In this system, 1 represents a rating of unreliable, while 5 represents a rating of very reliable.

Price ($)	Reliability	Price ($)	Reliability
12,000	1.4	15,500	2.8
12,000	4.2	15,500	3.6
20,500	3.2	51,000	4.1
20,500	3.7	51,000	2.2
17,000	3.1	18,500	3.5

a. Use this data to create a scatterplot of reliability versus price and determine the corresponding regression line.

b. Write a paragraph describing whether or not it is possible to use price to predict the reliability of a car. Use coefficients of determination and graphs to support your opinion.

* * * * * * * * * *

Summary Assessment

1. The following table shows the estimated population for 10 U.S. states, along with the amount of federal funding each state received during fiscal year 1995.

State	Estimated Population (millions)	Federal Funding (billions of dollars)
Alabama	4.3	22.8
Colorado	3.7	19.2
Florida	14.2	75.0
Indiana	5.8	23.0
Maine	1.2	6.6
Mississippi	2.7	14.3
Nebraska	1.6	7.5
North Dakota	0.6	3.8
Tennessee	5.3	26.6
Utah	2.0	8.6

Source: U.S. Bureau of the Census, 1996.

a. Create a scatterplot of federal funding received versus estimated population.

b. Find the equation of the regression line for the data.

c. How much of the variation in federal funding appears to be explained by population? Explain your response.

d. In 1995, Texas' estimated population was 18.7 million, while Minnesota's was 4.6 million. Use approximation intervals to predict a range for the federal funding received by each state.

e. In fiscal year 1995, Texas received $83.9 billion in federal funding, while Minnesota received $19.0 billion. How do these values compare with your predictions?

2. From a sample of at least 5 different people, collect data on the number of heartbeats for each of the intervals given in the following table.

5 sec	10 sec	15 sec	20 sec	25 sec

a. Create a graph of number of heartbeats versus time, along with the corresponding regression equation.

b. Describe how well your model explains any variation in the data.

c. What other factors besides time might influence number of heartbeats?

d. Use your model to predict the number of heartbeats a person might expect in 2.5 hr. Do you think your prediction is valid? Explain your response.

Module Summary

- The equation of the **mean line** for a set of data points in the form (x,y) is $y = \bar{y}$, where $\bar{y}$ represents the mean of the *y*-values.
- The **total variation** for a data set is the sum of the squares of the deviations from the mean line of the data points.
- The **explained variation** for a data set is the sum of the squares of the deviations from the mean line of the values predicted by the linear regression equation.
- The **coefficient of determination (r^2)** is the percentage of the total variation from the mean line explained by the linear regression equation, or:

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

  The value of r^2 is often reported as a decimal. It represents the proportion of the total variation in the *y*-values that can be explained by the linear relationship in the data set.
- The statistical association between two variables is referred to as **correlation**.
- The **linear correlation coefficient (*r*)** is found by taking the square root of the coefficient of determination r^2. The value of r is in the interval $[-1,1]$, where $|r| = \sqrt{r^2}$. The slope of the linear regression model determines if r is positive, negative, or zero.

  The closer $|r|$ is to 1, the stronger the linear relationship between the variables. When $r = 0$, the *y*-values in a data set are said to have no linear correlation to the *x*-values.
- A **residual** is the difference between the *y*-value of a data point and the corresponding *y*-value of the model.
- The **average deviation of prediction** is the square root of the average of the squared residuals. It can be found using the following formula, where *n* is the number of data points:

$$\text{average deviation of prediction} = \sqrt{\frac{\text{sum of the squares of the residuals}}{n}}$$

- As a rule of thumb, it is likely that the *y*-values of most data points with *x*-values near the mean ($\bar{x}$) fall within 2 average deviations of the regression line. Thus, it is likely that the predicted *y*-value will be within 2 average deviations of the true *y*-value. The interval of 2 average deviations of prediction on either side of the predicted *y*-value is an **approximation interval**.

Selected References

Consortium for Mathematics and Its Applications (COMAP). *Against All Odds: Inside Statistics*, Program 9. "Correlation." 30 min. 1989. Distributed by The Annenberg/CPB Collection, South Burlington, VT. Videocassette.

Meyers, R. A., ed. *Encyclopedia of Astronomy and Astrophysics*. San Diego, CA: Academic Press, 1989.

Swan, T. "Comparison Test: Luxury Sports Sedans." *Popular Mechanics*. (February 1993): 38–42.

U.S. Bureau of the Census. *Statistical Abstract of the United States 1993*. Washington, DC: U.S. Government Printing Office, 1994.

——. *Statistical Abstract of the United States 1995*. Washington, DC: U.S. Government Printing Office, 1996.

——. *Statistical Abstract of the United States 1996*. Washington, DC: U.S. Government Printing Office, 1997.