Linear Models

GUIDED SKILL SUPPORT

Overview

A trend line is a statistical tool that can model an association between two variables on a scatter plot. In Section 6.1, trend lines are sketched by hand.

Once the trend line is established, its slope and y-intercept can be used to solve problems about the association. In statistical analyses, the slope is often written as the expected change in the dependent variable (Δy) for a one-unit change in the independent variable ($\Delta x = 1$). The y-intercept is the predicted value of the dependent variable when the independent variable is zero. In scatter plots, the axes are often not drawn at the origin, so the y-intercept may not correspond to the point where the trend line crosses the vertical axis of the scatter plot.

Associations between two variables are described by their form, direction, strength, and outliers. Form refers to the shape of the pattern in the scatter plot: linear, some type of curve, or perhaps no pattern at all. Direction refers to whether the pattern is in general increasing (from left to right) or decreasing. For linear relationships, the slope can be used to describe both the steepness and the direction. The strength of the relationship indicates how much scatter there is in the data away from the trend line or model. Outliers are data points far removed from the bulk of the data.

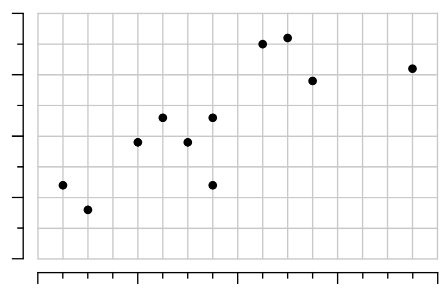
For additional information, see the Methods & Meanings boxes in Lessons 4.1.1, 4.2.1, and 6.1.2.

Example

1. Suppose you wanted to investigate the relationship between a student's GPA on a weighted 4-point scale and the amount of time they study in hours per week. GPA and study time data for 12 randomly-selected students are shown.

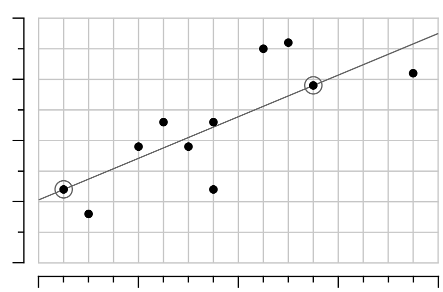
Study Time (hours/week)	4	5	11	1	15	2	10	6	7	0	7	9
GPA	2.9	3.3	3.9	2.2	4.1	1.8	4.6	2.9	2.2	3.0	3.3	4.5

a. Make a scatter plot without using technology.



b. Draw a trend line with a straightedge and determine its equation.

If it is reasonable, draw the trend line through a lattice point (where grid lines intersect) or a data point. That way, the equation will be easier to write. In this problem, the data points (1, 2.2) and (11, 3.9) appear to follow the trend in the data.



Using those two points, the equation of the trend line is $y = 2.03 + 0.17x$, where y is the predicted GPA and x is the time the student studies (in hours per week).

Example continues on next page.

1. Example continued from previous page.

c. Interpret the slope and y-intercept in terms of study time and GPA.

The slope indicates that a student's GPA is expected to increase by 0.17 points for every additional hour of studying per week. The y-intercept predicts that students who do not study at all have a GPA of 2.03.

d. Describe the association.

To describe an association means to describe the form, direction, strength, and outliers. The form appears linear: it does not appear curved nor like a collection of randomly scattered points. The direction is positive: students who study more tend to have higher GPAs. A student's GPA is expected to increase by 0.17 points for every additional hour of studying per week. The strength is moderate: it is strong enough to easily see its form, but there is scatter about the line. There do not appear to be any outliers.

In summary, there is a moderately strong linear relationship between study time and GPA for students, with no outliers in the data.

Practice

1. Suppose a random sample of 10 car models is selected and the engine power (horsepower, hp) and mileage (miles per gallon, mpg) are recorded for each.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15

a. Create a scatter plot of the data.

b. Estimate a trend line and determine its equation.

c. Interpret the slope and y-intercept in the context of engine power and mileage.

d. Describe the association.

2. The principal at University High School went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. Their final exam scores for each course are in the following table.

AP Music Score	88	74	82	64	97	90	82	72	78	62
AP English Score	63	96	86	90	68	90	78	74	96	79

a. Create a scatter plot of the data.

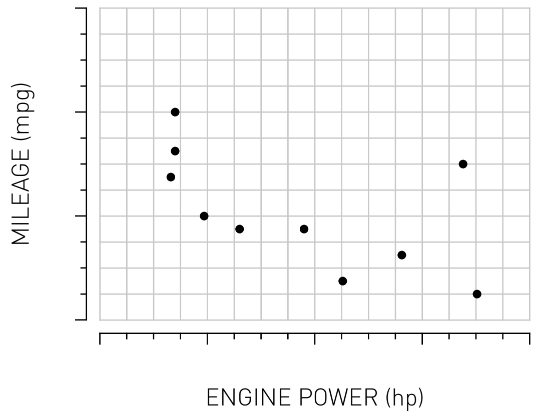
b. Estimate a trend line and determine its equation.

c. Interpret the slope and y-intercept in context.

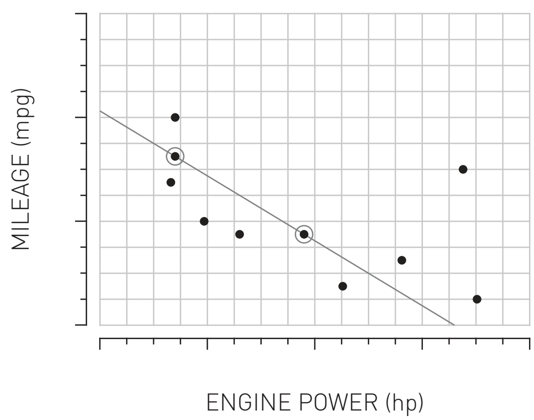
d. Describe the association.

Answers

1. a.



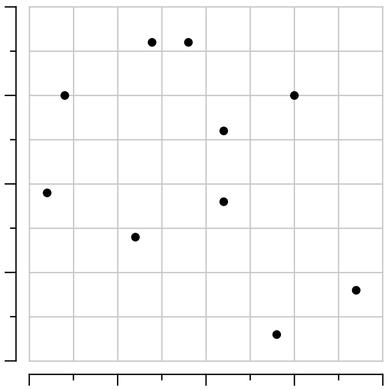
b. Responses may vary. If you ignore the possible outlier at (438, 20), a reasonable trend line could pass through the points (170, 21) and (290, 15). Using those points, the equation of the trend line is $y = 29.5 - 0.05x$.



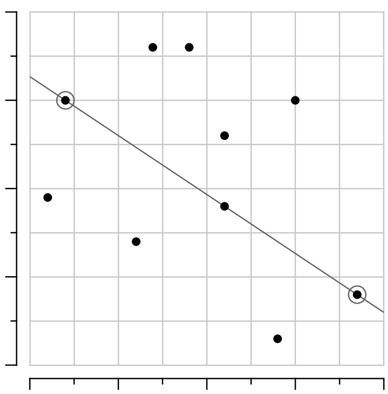
c. For every increase of 1 horsepower, the gas mileage is expected to decrease by 0.05 mpg. The *y*-intercept means that a 0-horsepower car would get 29.5 mpg, which does not make sense. The *y*-axis is far outside of the data, so it represents an extrapolation. Prediction models are unreliable when you extrapolate them.

d. When the outlier at (438, 20) is removed, there appears to be a strong, negative, linear relationship. If the outlier is not removed, the association is more moderate (not as strong).

2. a.



b. Responses may vary. A reasonable trend line could pass through the points (64, 90) and (97, 68). Using these points, the equation of the trend line is $y = 132.7 - 0.67x$.



c. An increase of 1 point for the music score results in a predicted decrease of 0.67 points for the English score. The y-intercept would mean that a student who scored 0 on the music test is expected to score 133 on the English test. This does not make sense in the given context. The y-axis is far outside of the data, so this prediction represents an extrapolation. Prediction models are unreliable when you extrapolate them.

d. There is a lot of scatter around the trend line. The linear association is weak. There are no apparent outliers.

Residuals and Upper and Lower Bounds

GUIDED SKILL SUPPORT

Overview

A **residual** is a measure of how far a prediction is from what is actually observed. Residuals are calculated in the vertical (*y*) direction between a data point and the corresponding point on the graph of the model.

residual = actual – predicted

A positive residual means that the actual value is greater than the predicted value; a negative residual means that the actual value is less than the predicted value.

A prediction made from a trend line gives no indication of the variability in the original data. Upper and lower boundary lines are parallel lines above and below the trend line. They give upper and lower limits (a "margin of error") to predictions made by the trend line. For example, predicting a test score of 87 is useful. But a prediction of 87 with a lower bound of 86 and an upper bound of 88 is very different from one with a lower bound of 77 and an upper bound of 97. The upper and lower boundary lines help set limits on predictions.

The most commonly used techniques for setting upper and lower boundary lines are beyond the scope of this course. In this course, the bounds represent the observable range in the data and are established by determining the residual with the greatest distance, and then adding and subtracting that distance from the prediction.

For additional information, see the Methods & Meanings box in Lesson 6.1.4. For additional examples and more practice on the topics from this chapter, see the Checkpoint 8 materials.

Example

1. Suppose you wanted to investigate the relationship between a student's GPA on a weighted 4-point scale and the amount of time they study in hours per week. GPA and study time data for 12 randomly-selected students are shown. An estimated trend line for this data could be $y = 2.03 + 0.17x$, where x is the study time (in hours per week) and y is the predicted GPA.

Study Time (hours/week)	4	5	11	1	15	2	10	6	7	0	7	9
GPA	2.9	3.3	3.9	2.2	4.1	1.8	4.6	2.9	2.2	3.0	3.3	4.5

a. Calculate the residual for the student who studied 10 hours per week, and interpret it in context. Would a student prefer a positive or a negative residual?

From the trend line, a student who studies 10 hours per week is predicted to earn a GPA of $y = 2.03 + 0.17(10) = 3.73$. But this student actually received a 4.6 GPA. The residual is:

$$\text{residual} = \text{actual} - \text{predicted}$$
$$\text{residual} = 4.6 - 3.73$$
$$\text{residual} = 0.87$$

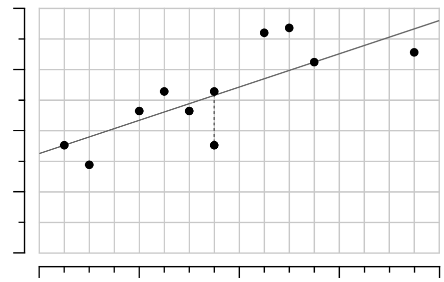
The student who studied 10 hours per week earned a GPA that was 0.87 points higher than predicted. A student would prefer a positive residual because it means that they earned a higher GPA than was predicted from the time spent studying.

Example continues on next page.

1. Example continued from previous page.

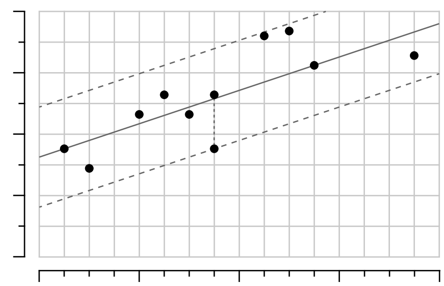
b. Write the equations for the upper and lower boundary lines, and show the boundary lines on a scatter plot.

The upper and lower bounds can be calculated using the data point that has the largest residual. The largest residual (the largest vertical distance from the trend line) is at the point (7, 2.2). The predicted GPA for a student who studies 7 hours is $y = 2.03 + 0.17(7) = 3.22$. The residual is:



residual = actual – predicted
residual = 2.2 – 3.22
residual = –1.02

The upper and lower bounds are lines parallel to the trend line. The upper bound is 1.02 units above the trend line, and the lower bound is 1.02 units below the trend line.



The upper boundary line is: $y = 3.05 + 0.17x$.
The lower boundary line is: $y = 1.01 + 0.17x$.

c. Predict the upper and lower bounds of the GPA for a student who studies 8 hours per week. Is your prediction useful?

The predicted lower bound for a student who studies 8 hours per week is $y = 1.01 + 0.17(8) = 2.37$, and the predicted upper bound is $y = 3.05 + 0.17(8) = 4.41$. A student who studies 8 hours per week is predicted to have a GPA between 2.37 and 4.41. Due to the large amount of variability in the data collected, this is a large range. The prediction is not particularly useful.

Practice

1. Suppose a random sample of nine car models is selected and the engine power (horsepower, hp) and mileage (miles per gallon, mpg) are recorded for each. A trend line was estimated to be $y = 29.5 - 0.05x$, where y is the predicted mileage and x is the engine power.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15

a. Calculate the residual for the vehicle that has 381 horsepower, and interpret it in context. Would a car owner prefer a positive or negative residual?

b. Write the equations for the upper and lower boundary lines, and draw the boundary lines on a scatter plot.

c. Predict the upper and lower bounds of the gas mileage for a car with 300 horsepower. Is your prediction useful?

2. The principal at University High School went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. Their final exam scores for each course are in the following table. A trend line for this data was estimated to be $y = 132.7 - 0.67x$, where y is the predicted AP English score and x is the AP Music score.

AP Music Score	88	74	82	64	97	90	82	72	78	62
AP English Score	63	96	86	90	68	90	78	74	96	79

a. Write the equations for the upper and lower boundary lines, and draw the boundary lines on a scatter plot.

b. Predict the upper and lower bounds of the AP English score for a student with a perfect score of 100 on the music final. Is your prediction useful?

Answers

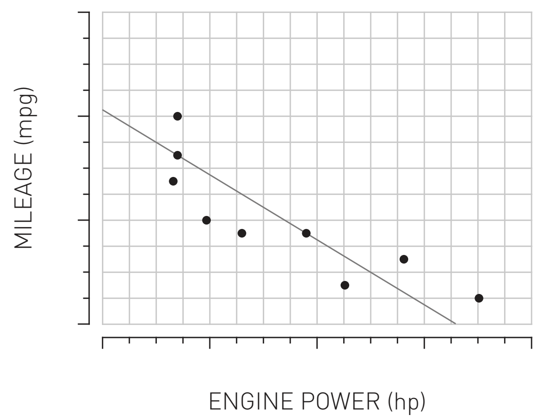
1. a. From the trend line, a vehicle with 381 horsepower is predicted to have a gas mileage of $y = 29.5 - 0.05(381) \approx 10.5$ mpg. But this vehicle actually got 13 mpg. The residual is:

$$\begin{aligned}\text{residual} &= \text{actual} - \text{predicted}\\ \text{residual} &= 13 - 10.5\\ \text{residual} &= 2.5\end{aligned}$$

The vehicle with 381 horsepower got 2.5 mpg more than was predicted. A car owner would prefer a positive residual—a positive residual means the car is getting more miles per gallon than was predicted from its power.

b. Without the outlier, the largest residual (the largest vertical distance from the trend line) seems to be at the point (197, 16). The predicted mileage for a car with 197 horsepower is $y = 29.5 - 0.05(197) = 19.7$ mpg. The residual for this point is:

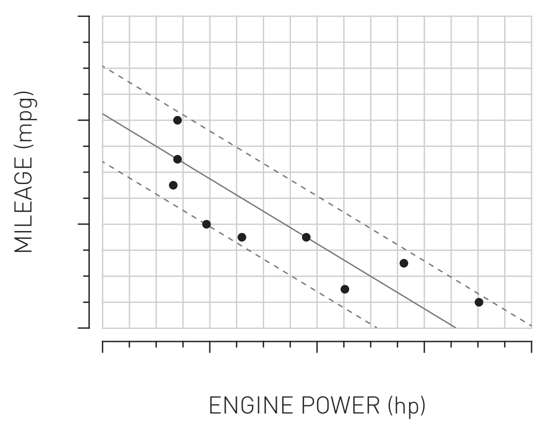
$$\begin{aligned}\text{residual} &= \text{actual} - \text{predicted}\\ \text{residual} &= 16 - 19.7\\ \text{residual} &= -3.7\end{aligned}$$



The upper and lower bounds are lines parallel to the trend line. The upper bound is 3.7 units above the trend line, and the lower bound is 3.7 units below the trend line.

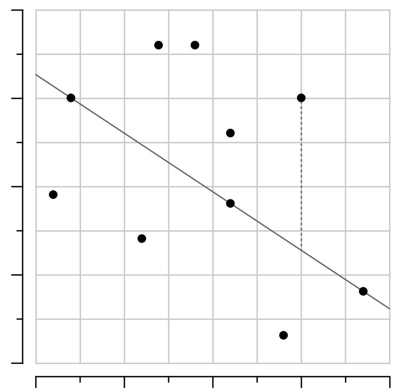
The upper boundary line is: $y = 33.2 - 0.05x$.
The lower boundary line is: $y = 25.8 - 0.05x$.

c. The upper bound is $y = 33.2 - 0.05(300) = 18.2$, and the lower bound is $y = 25.8 - 0.05(300) = 10.8$. A car with 300 horsepower is predicted to have a gas mileage between 10.8 and 18.2 miles per gallon. The prediction could be useful: it indicates that this is a car with low gas mileage.



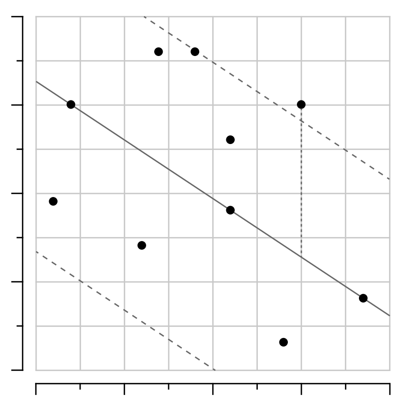
2. a. The largest residual (the largest vertical distance from the trend line) is at the point (90, 90). The predicted AP English exam score for a student with a score of 90 on the AP Music exam is $y = 132.7 - 0.67(90) = 72.4$. The residual for this point is:

residual = actual – predicted
residual = 90 – 72.4
residual = 17.6



The upper boundary line is: $y = 150.3 - 0.67x$. The lower boundary line is: $y = 115.1 - 0.67x$.

b. The upper bound is $y = 150.3 - 0.67(100) = 83.3$, and the lower bound is $y = 115.1 - 0.67(100) = 48.1$. A student who scores 100 on the AP Music final exam is predicted to earn a score between 48 and 83 on the AP English final exam. This range is large and the context may involve more variables than exam scores, so the prediction is not very useful.



Least-Squares Regression Line

GUIDED SKILL SUPPORT

Overview

A **least-squares regression line** (LSRL) of scattered data is a unique line that minimizes the sum of the squared residuals. The LSRL is sometimes called a **line of best fit**.

For additional information, see the Methods & Meanings box in Lesson 6.2.1, and for additional examples and more practice on the topics from this chapter, see Checkpoint 8.

Example

1. Suppose you wanted to investigate the relationship between a student's GPA on a weighted 4-point scale and the amount of time they study in hours per week. GPA and study time data for 12 randomly-selected students are shown.

Study Time (hours/week)	GPA
4	2.9
5	3.3
11	3.9
1	2.2
15	4.1
2	1.8
10	4.6
6	2.9
7	2.2
0	3.0
7	3.3
9	4.5

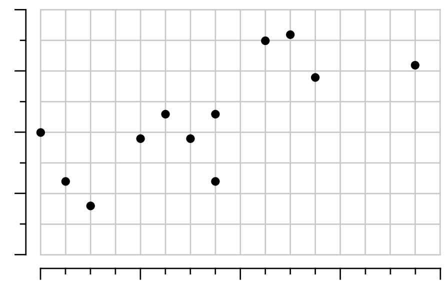
a. Determine the equation of the LSRL.

> It is important to define the variables and units when creating an LSRL equation.
>
> Let x be the number of hours spent studying each week, and y be the predicted GPA for that student. The LSRL is $y = 2.249 + 0.152x$.

Example continues on next page.

1. Example continued from previous page.

b. Use graphing technology to make a scatter plot of the data. Graph the LSRL on the scatter plot.



c. Calculate the residuals of the data in the table. What is the residual for the student who spent 10 hours per week studying? Interpret the residual in context.

The residual for the student who studied 10 hours per week is 0.83 grade points. That means the student earned a GPA that was 0.83 grade points higher than was predicted by the least-squares regression line.

Study Time (hours/week)	GPA	Residual
4	2.9	0.043
5	3.3	0.291
11	3.9	−0.021
1	2.2	−0.021
15	4.1	−0.429
2	1.8	−0.753
10	4.6	0.831
6	2.9	−0.261
7	2.2	−1.113
0	3.0	0.751
7	3.3	−0.013
9	4.5	0.883

Practice

1. Suppose a random sample of nine car models is selected and the engine power (horsepower, hp) and mileage (miles per gallon, mpg) are recorded for each.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15

a. Determine the equation of the LSRL. What do you predict the gas mileage will be for a car with 400 horsepower?

b. Use graphing technology to make a scatter plot of the data. Graph the LSRL on the scatter plot.

c. Calculate the residuals of the data in the table. What is the residual for the 230-horsepower car? Interpret the residual in context.

2. The principal at University High School went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. Their final exam scores for each course are in the following table.

AP Music Score	AP English Score
88	63
74	96
82	86
64	90
97	68
90	90
82	78
72	74
78	96
62	79

a. Determine the equation of the LSRL.

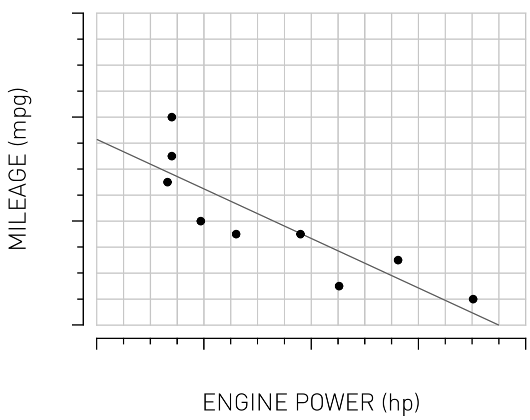
b. Use graphing technology to make a scatter plot of the data. Graph the LSRL on the scatter plot.

c. Calculate the residuals of the data in the table. Would a student prefer a positive or a negative residual in this situation?

Answers

1. a. Let y be the predicted gas mileage (in miles per gallon), and x be the engine power (in horsepower). The LSRL is $y = 26.11 - 0.0382x$. A car with 400 horsepower is predicted to have gas mileage of $26.11 - 0.0382(400) = 10.8$ mpg.

b.



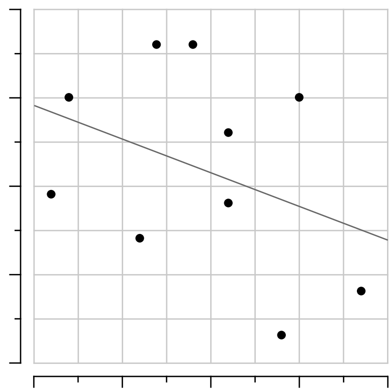
c.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15
Residual	−2.585	4.384	−0.769	−2.324	1.444	1.384	−2.657	1.118	−0.032

The car with 230 horsepower has a residual of −2.3 mpg. That means the car has a gas mileage that is 2.3 mpg less than what would be predicted by the LSRL.

2. a. Let y be the predicted AP English final exam score, and x be the AP Music final exam score. The LSRL is $y = 112.1 - 0.3813x$.

b.



c.

AP Music Score	AP English Score	Residual
88	63	−15.55
74	96	12.12
82	86	5.17
64	90	2.30
97	68	−7.11
90	90	12.22
82	78	−2.83
72	74	−10.65
78	96	13.64
62	79	−9.46

A student would prefer a positive residual. That would mean their AP English score is higher than the LSRL predicted.

Residual Plots and Correlation

GUIDED SKILL SUPPORT

Overview

A **residual plot** is created to analyze the appropriateness of a linear model. See the Methods & Meanings box in Lesson 6.2.2 for additional support with residual plots.

The **correlation coefficient**, *r*, is a measure of how much or how little data is scattered around the LSRL; it is a measure of the strength of a linear association. See the Methods & Meanings box in Lesson 6.2.3 for additional support with the correlation coefficient.

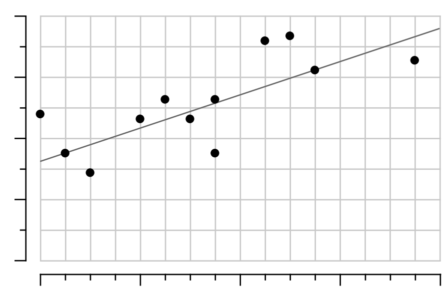
For additional examples and more practice on the topics from this chapter, see Checkpoint 8.

Example

1. Suppose you wanted to investigate the relationship between a student's GPA on a weighted 4-point scale and the amount of time they study in hours per week. GPA and study time data for 12 randomly-selected students are shown.

Study Time (hours/week)	4	5	11	1	15	2	10	6	7	0	7	9
GPA	2.9	3.3	3.9	2.2	4.1	1.8	4.6	2.9	2.2	3.0	3.3	4.5

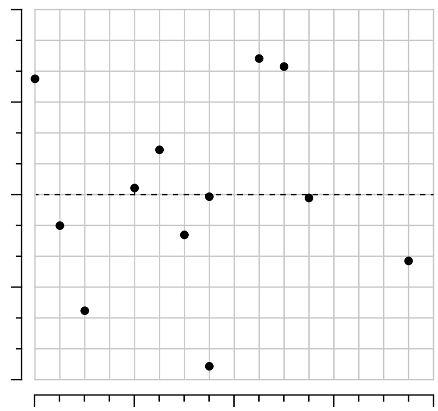
a. Graph the LSRL on a scatter plot of the data.



Example continues on next page.

1. Example continued from previous page.

b. Create and interpret a residual plot.



There is no apparent pattern in the residual plot—it looks like randomly scattered points—which means a linear model (instead of a curve) is the most appropriate way to model the relationship.

c. Determine the correlation coefficient, r. Is the association strong or weak?

$r \approx 0.7338$; The association is strong.

d. Fully describe the association between GPA and hours studied.

When describing an association, the form, direction, strength, and outliers should be included. The least-squares regression line is $y = 2.249 + 0.152x$, where y is the predicted GPA, and x is the study time in hours per week.

The form in the scatter plot appears to be linear; it does not appear to be curved nor simply a collection of randomly-scattered points. The residual plot shows random scatter, confirming that a linear model is an appropriate choice.

The direction is positive; as students study more, their GPAs are typically higher. From the slope, a student's GPA is expected to increase by 0.15 points for every additional hour of studying per week.

The correlation coefficient is 0.73, confirming a strong linear association. There are no apparent outliers.

Practice

1. Suppose a random sample of nine car models is selected and the engine power (horsepower, hp) and mileage (miles per gallon, mpg) are recorded for each.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15

a. Graph the LSRL on a scatter plot of the data.

b. Create and interpret a residual plot.

c. Determine the correlation coefficient, *r*. Is the association strong or weak?

d. Fully describe the association between mileage and engine power.

2. The principal at University High School went through the records for the past year and found 10 students who were enrolled in both Advanced Placement Music and Advanced Placement English. Their final exam scores for each course are in the following table.

AP Music Score	88	74	82	64	97	90	82	72	78	62
AP English Score	63	96	86	90	68	90	78	74	96	79

a. Graph the LSRL on a scatter plot of the data.

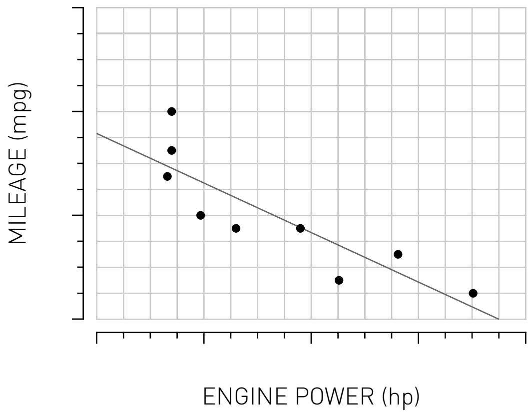
b. Create and interpret a residual plot.

c. Determine the correlation coefficient, *r*. Is the association strong or weak?

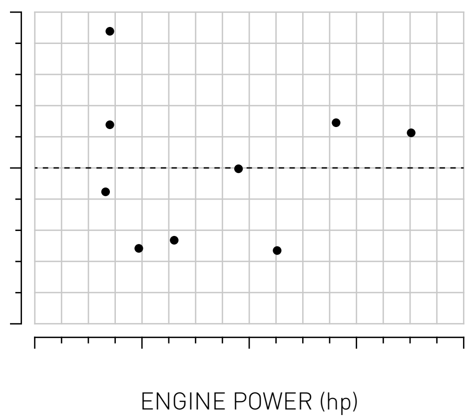
d. Fully describe the association between AP English scores and AP Music scores.

Answers

1. a.



b.



There is an apparent U-shaped pattern in the residual plot, which suggests a curved model would be the most appropriate way to model the relationship, not a line. Nonetheless, by visual examination of the scatter plot and LSRL, a linear model is not completely inappropriate.

c. $r \approx -0.8602$; The association is strong.

d. The least-squares regression line is $y = 26.11 - 0.0382x$, where y is the predicted mileage (in mpg) and x is the engine power (in hp).

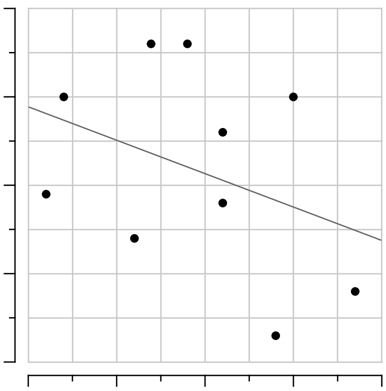
The residual plot indicates a curved model might fit the data better than a linear model, but by observation of the scatter plot, the LSRL models the data well enough to proceed. Therefore, the form is linear.

The direction is negative; the model predicts that for every increase of 1 horsepower, the mileage is expected to decrease by 0.0382 mpg.

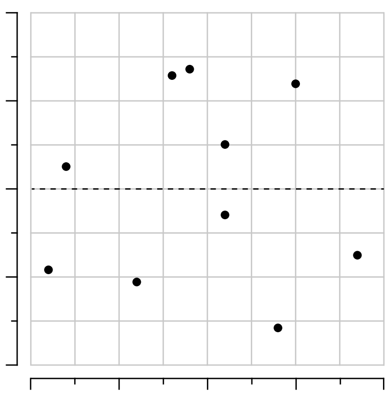
The association's strength is strong, with a correlation coefficient of –0.8602.

There are no apparent outliers.

2. a.



b.



There is no apparent pattern in the residual plot—it looks like randomly scattered points—which means a linear model (instead of a curve) is the most appropriate way to model the relationship.

c. $r \approx -0.3733$; The association is moderate.

d. The LSRL is $y = 112.1 - 0.3813x$, where y is the predicted AP English score, and x is the AP Music score.

The residual plot confirms that a linear model is appropriate. Therefore, the form is linear.

The association is negative; the model predicts that for every increase of 1 point in the AP Music score, the AP English score will drop by 0.38 points.

The association is moderate, with a correlation coefficient of approximately –0.3733.

There are no apparent outliers.

Association and Causation

GUIDED SKILL SUPPORT

Overview

Considering all of the statistical tools available to measure the correlation between two variables, it is tempting to believe that a strong **association** between two variables is evidence of **causation**. However, the fact that two variables move together in a predictable manner does not show that one causes the other. Cause can usually only be shown through carefully controlled experiments, and not by just observing the relationship between variables.

Example

1. A city is proposing a new bus route with several added bus stops. A community member who lives along the newly-proposed route does some research, which shows a strong association between the number of bus stops and the air pollution of a community.

Assume that the association this community member is presenting is true. Suggest a hidden or lurking variable that may be the actual cause of the relationship.

> While it may be true that communities with more bus stops have higher pollution, it is unlikely that the number of bus stops is causing the pollution. The lurking variable may be the number or frequency of buses visiting the bus stops. Increased pollution may also result if the bus stops are located in a densely populated community with more automobile traffic in general or with more industries that pollute the air.

Practice

1. For each fake headline, assume that the association is true. Suggest a hidden or lurking variable that may be the actual cause of the relationship.

a. “BOOKS IN THE HOME ASSOCIATED WITH HIGHER TEST SCORES”

“Students buying truckloads of books before taking national exams.”

b. “CHESS CLUB BOASTS HIGHEST IVY LEAGUE ADMISSION RATES”

“Club president now flooded with membership applications.”

c. “STUDENTS WHO LIVE OFF CAMPUS AT HIGHER RISK FOR AUTO ACCIDENTS”

“State Police Warn: Play it safe, live on campus.”

Answers

1. a. A person who does a lot of reading would likely have a larger number of books in their home and higher test scores. So reading regularly is likely the cause. Just buying or possessing books may not cause higher test scores.

b. Chess is a game of strategy and an extracurricular activity. Those who play regularly may be more inclined to apply to Ivy League schools than non-chess players. Just being in a chess club may not cause greater Ivy League school admission rates.

c. Students who live off campus probably commute to school and may spend more time driving to and from school than students who live on campus. The additional hours on the road may be the cause of more accidents.

Nonlinear Models

GUIDED SKILL SUPPORT

Overview

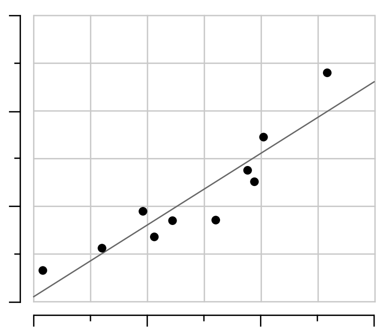
Many relationships are modeled better by a curve than by a line. For additional information, see the Methods & Meanings box in Lesson 6.3.3.

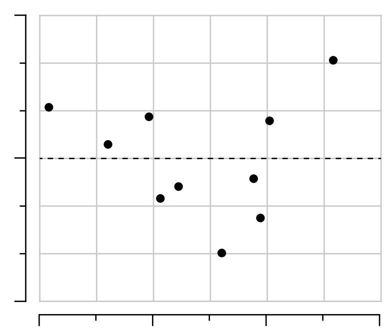
Examples

1. Methyl mercury is a neurotoxin found in many fish species. Plankton (plant-like organisms in water) absorb mercury from contaminated runoff water, then small fish eat the plankton, and the methyl mercury travels up the food web to larger predators. In general, the larger the fish, the higher the concentration of methyl mercury in its flesh. Assume a biologist is sampling white bass from a lake known to have mercury in its water in an effort to find a relationship between the length of the fish and the concentration of methyl mercury in the fish. The following data was collected from a sample of 10 white bass (ppm is "parts per million").

Length (cm)	24.4	23.0	21.1	24.7	25.1	20.3	18.0	27.9	19.8	15.4
Mercury (ppm)	0.689	0.428	0.425	0.629	0.863	0.340	0.281	1.20	0.474	0.164

The LSRL was calculated, and the following scatter plot and residual plot were made.





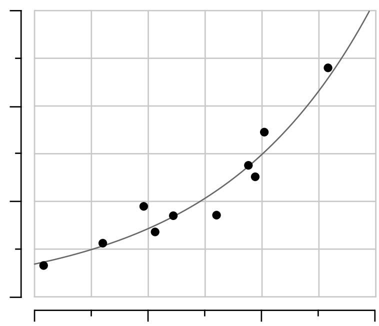
The curved pattern in the residual plot indicates that a nonlinear model would be a better fit.

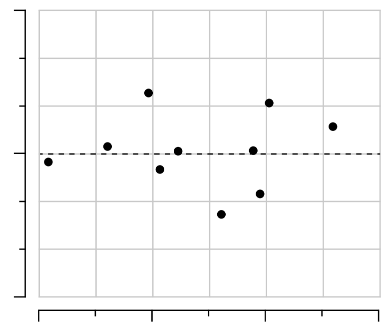
Example continues on next page.

1. Example continued from previous page.

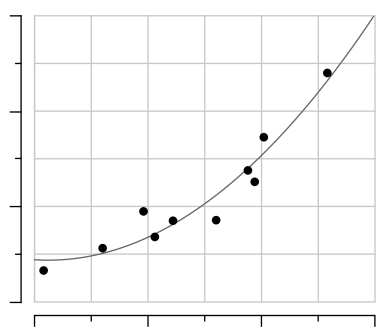
a. Perform exponential and quadratic model regressions for the data and make the associated scatter and residual plots.

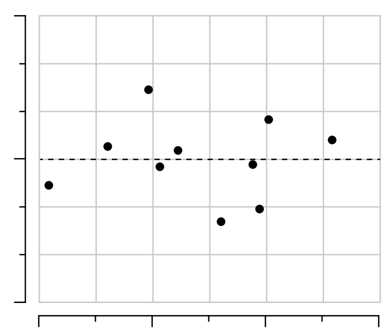
EXPONENTIAL MODEL





QUADRATIC MODEL





b. Use the residual plots to select the best nonlinear model.

Try to identify the model with the least patterning or the most "random-looking" scatter.

Both the exponential and quadratic residual plots look fairly random; either model would be an appropriate choice based on the plots. But successive growth like this—in this case, the buildup of mercury as the fish ages and thus gets longer—is more likely explained by an exponential relationship than a quadratic one. The exponential model is a good choice to describe the association between the length of the fish and the concentration of mercury.

Example continues on next page.

1. Example continued from previous page.

c. Use your model to predict the amount of mercury in a 20 cm fish.

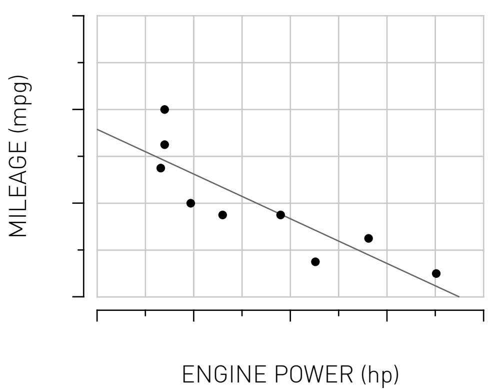
The exponential regression is $y = 0.0187 \cdot 1.1588^x$, where x is the length of the fish in cm and y is the predicted concentration of mercury in parts per million. If $x = 20$ cm, then $y = 0.356$. A 20 cm fish is expected to contain 0.356 ppm of mercury.

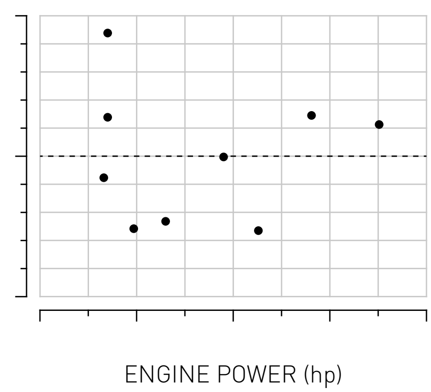
Practice

1. Suppose a random sample of nine car models is selected and the engine power (horsepower, hp) and mileage (miles per gallon, mpg) are recorded for each.

Engine Power (hp)	197	170	166	230	381	170	326	451	290
Mileage (mpg)	16	24	19	15	13	21	11	10	15

The LSRL was calculated, and the following scatter plot and residual plot were made.





The curved pattern in the residual plot indicates that a nonlinear model would be a better fit.

a. Perform exponential and quadratic regressions for the data and make the associated scatter and residual plots.

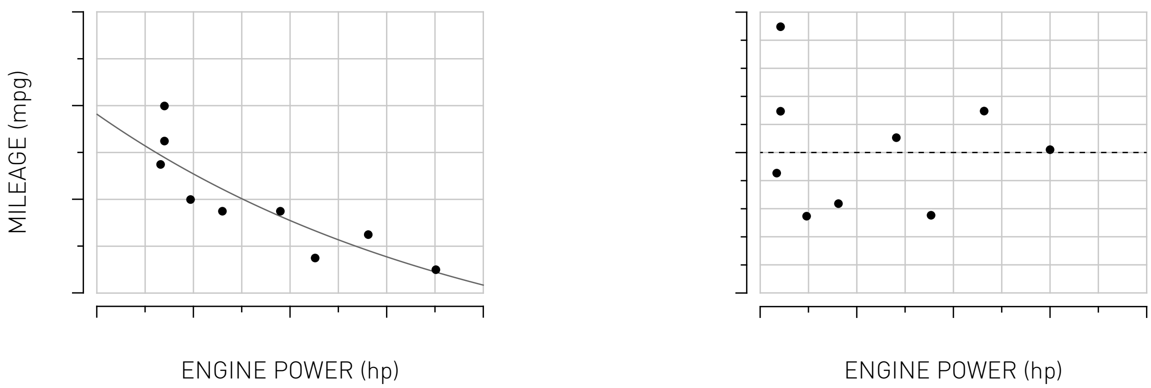
b. Use the residual plots to select the best nonlinear model.

c. What do you predict the gas mileage will be for a car with 400 horsepower?

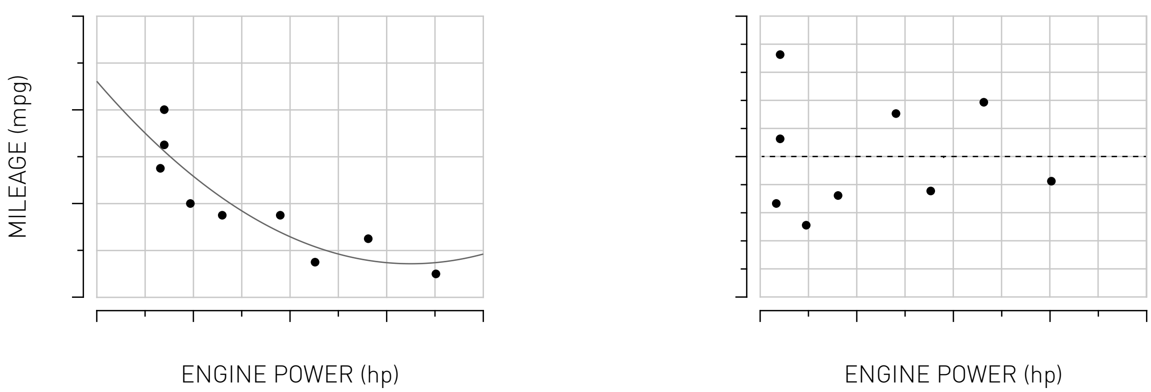
Answers

1. a.

EXPONENTIAL MODEL



QUADRATIC MODEL



b. Both exponential and quadratic residual plots appear equally scattered. However, the quadratic model turns back up, indicating that for higher engine power, the gas mileage increases. That does not make sense. Therefore, the exponential model is a better representation.

c. The exponential model equation is $y = 32.07882 \cdot 0.997239^x$, where x is the engine power (in hp) and y is the predicted gas mileage (in mpg). If $x = 400$, then $y = 11.03$. A 400-horsepower car is predicted to get about 11 miles per gallon.

