Chapter 4 Summarizing Bivariate Data
Sec 4.1 Correlation

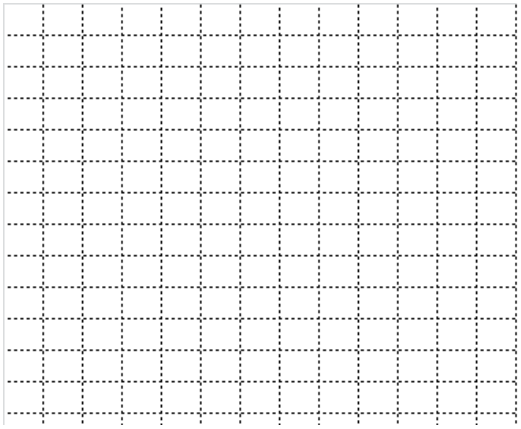
Objectives
1. Construct scatterplots for bivariate data
2. Compute the correlation coefficient
3. Interpret the correlation coefficient
4. Understand that correlation is not the same as causation

Objective 1: Construct scatterplots for bivariate data

A scatter plot is a graph in which the values of two variables are plotted along two axes. In a scatterplot each individual in the data set contributes an ordered pair of numbers.

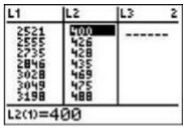
Example 1
The following table presents the size in square feet and the selling price in thousands of dollars, for a sample of houses in a suburban Denver neighborhood. Construct a scatterplot for the data.

Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

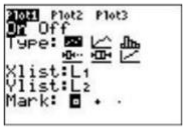


Scatterplots on the TI-84 PLUS

The following steps will create a scatterplot for the house sizes and prices data on the TI-84 PLUS.

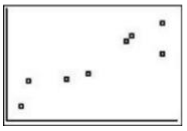


Step 1: Enter the x-values in **L1** and the y-values in **L2**.



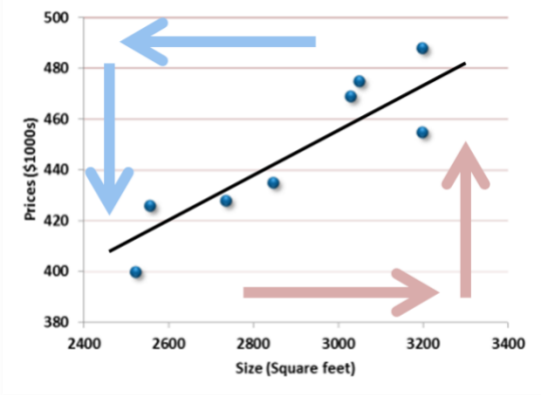
Step 2: Press **2nd,Y=**, then **1** to access the Plot1 menu. Select **On** and the scatterplot type.

Step 3: Press **Zoom, 9** to view the plot.



Positive Linear Association

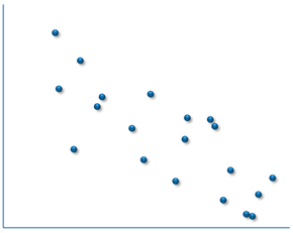
- Observe that larger sizes tend to be associated with larger prices, and smaller sizes tend to be associated with smaller prices. We refer to this as a **positive association** between size and selling price.
- In addition, the points tend to cluster around a straight line. We describe this by saying that the relationship between the two variables is **linear**.
- Therefore, we can say that the scatterplot exhibits a **positive linear association** between size and selling price.



Other Types of Association

Two variables are **positively associated** if large values of one variable are associated with large values of the other.

Two variables are **negatively associated** if large values of one variable are associated with small values of the other.



Two variables have a **linear relationship** if the data tend to cluster around a straight line when plotted on a scatterplot.

Objective 2: Compute the Correlation Coefficient

A numerical measure of the strength of the linear relationship between two variables is called the **correlation coefficient**.

Correlation Coefficient

Given ordered pairs (x, y) with sample means $\bar{x}$ and $\bar{y}$, sample standard deviations s_x and s_y and sample size n, the correlation coefficient r is given by

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Properties:

- The correlation coefficient is always between –1 and 1. That is, $-1 \le r \le 1$.
- The correlation coefficient does not depend on the units of the variables.
- It does not matter which variable is x and which is y.
- The correlation coefficient only measures the strength of the ***linear*** relationship.
- The correlation coefficient is **sensitive to outliers** and can be misleading when outliers are present.

Example 2a (Manual Method – Do not follow this method. This example is just for your reference.)

Compute the correlation between size and the selling price:

Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Step 1: Compute the sample means and standard deviations. We obtain $\bar{x}$= 2891.25, $\bar{y}$ = 447.0, s_x = 269.49357, s_y = 29.68405.

x	y	$\frac{x-\bar{x}}{s_x}$	$\frac{y-\bar{y}}{s_y}$	$\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
2521	400	−1.3738732	−1.5833419	2.1753110
2555	426	−1.2477106	−0.7074506	0.8826936
2735	428	−0.5797912	−0.6400744	0.3711095
2846	435	−0.1679075	−0.4042575	0.0678779
3028	469	0.5074333	0.7411387	0.3760785
3049	475	0.5853572	0.9432675	0.5521484
3198	488	1.1382461	1.3812131	1.5721604
3198	455	1.1382461	0.2695050	0.3067630

$$\frac{\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1} = \frac{6.3041423}{7} = 0.9005918$$

To Find Correlation Coefficient on the TI-84 PLUS

Before computing the least-squares regression line, a one-time calculator setting should be modified to correctly configure the calculator to display the correlation coefficient. The following steps describe how to do this.

Step 1: Press **2nd, 0** to access the calculator catalog.

Step 2: Scroll down and select **DiagnosticOn**.

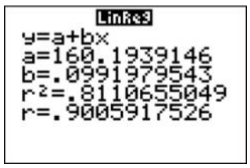
Step 3: Press **Enter** twice.

The following steps describe how to compute the least-squares regression line using the TI-84 PLUS calculator for the house size and selling price data.

Step 1: Enter the x -values in to **L1** and the y -values into **L2**.

Step 2: Press **STAT** and the right arrow key to access the **CALC** menu.

Step 3: Select the **LinReg(a+bx)** command. Verify that **L1** is entered in the **Xlist** field and **L2** in the **Ylist** field. Select **Calculate**.



Example 2b

Use your calculator to find the correlation coefficient

Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Objective 3: Interpreting the Correlation Coefficient

The correlation coefficient can be interpreted as follows:

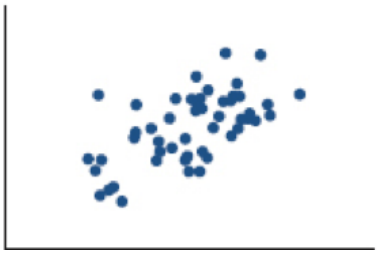
- If r is positive, the two variables have a positive linear association.
- If r is negative, the two variables have a negative linear association.
- If r is close to 0, the linear association is weak.
- The closer r is to 1, the more strongly positive the linear association is.
- The closer r is to −1, the more strongly negative the linear association is.
- If r = 1, then the points lie exactly on a straight line with positive slope; in other words, the variables have a perfect positive linear association.
- If r = −1, then the points lie exactly on a straight line with negative slope; in other words, the variables have a perfect negative linear association.

When two variables are **not linearly related**, the correlation coefficient does not provide a reliable description of the relationship between the variables.

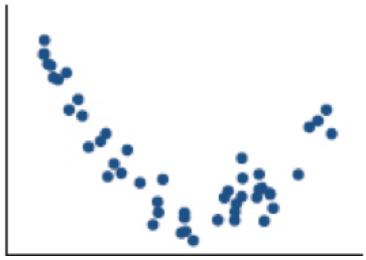
Example 3

Determine whether the correlation coefficient is an appropriate summary for the scatterplot.

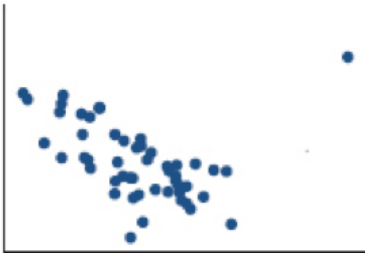
(a)



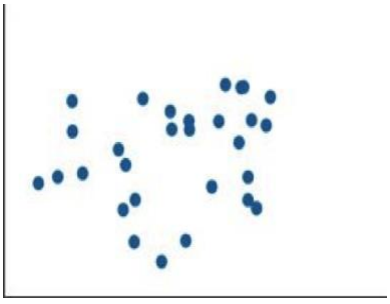
(b)



(c)



(d)



Objective 4: Understand that correlation is not the same as causation

A group of elementary school children took a vocabulary test. It turned out that children with larger shoe sizes tended to get higher scores on the test, and those with smaller shoe sizes tended to get lower scores.
As a result, there was a large positive correlation between vocabulary and shoe size.

Does this mean that learning new words causes one's feet to grow, or that growing feet cause one's vocabulary to increase?

The fact that shoe size and vocabulary are correlated does not mean that changing one variable will cause the other to change.

Correlation is not the same as causation. In general, when two variables are correlated, we cannot conclude that changing the value of one variable will cause a change in the value of the other.

Example 4
The following table presents the average price in dollars for a dozen eggs and a gallon of milk for each month from January through December 2012.

Dozen Eggs	1.94	1.80	1.77	1.83	1.69	1.67	1.65	1.88	1.89	1.96	1.96	2.01
Gallon of Milk	3.58	3.52	3.50	3.47	3.43	3.40	3.43	3.47	3.47	3.52	3.54	3.58

The correlation coefficient between the price of eggs and the price of milk is found to be 0.845

(a) In a month where the price of eggs is above average, would you expect the price of milk to be above average or below average?

(b) Which of the following is the best interpretation of the correlation coefficient? Select all the answers that apply.

(i) When the price of eggs rises, it causes the price of milk to rise.

(ii) When the price of milk rises, it causes the price of eggs to rise,

(iii) Changes in the price of eggs or milk do not cause the changes in the price of the other.

(iv) The correlation indicates that the prices of milk and eggs tend to go up and down together.

Sec 4.2 The Least-Squares Regression Line

Objectives

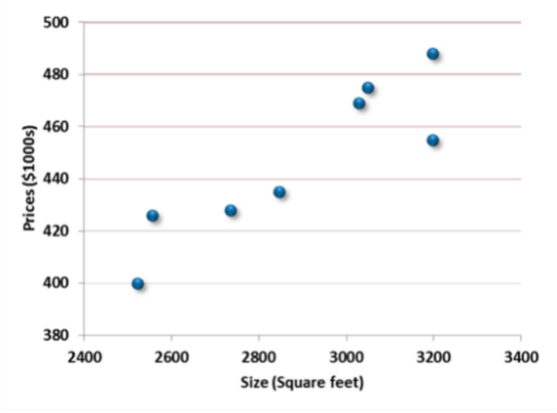
1. Compute the least-squares regression line
2. Use the least-squares regression line to make predictions
3. Interpret predicted values, the slope, and the y-intercept of the least-squares regression line

Objective 1: Compute the least-squares regression line

Example 1

The following table presents the size in square feet and the selling price in thousands of dollars, for a sample of houses in a suburban Denver neighborhood. Construct a scatterplot for the data.

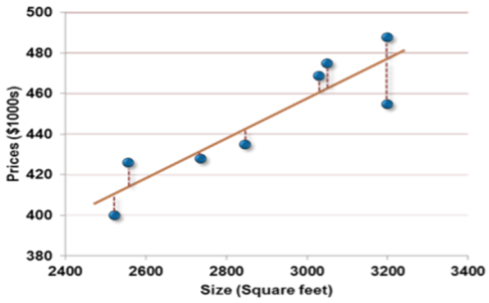
Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

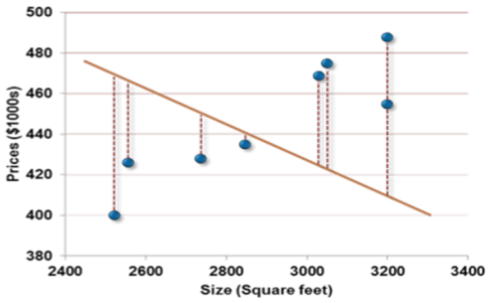


We concluded that there is a **strong positive linear association** between size and sales price. We can use these data to predict the selling price of a house based on its size.

Least-Square Regression Line

The figures present scatterplots of the above data, each with a different line superimposed. It is clear that the line in the figure on the left fits better than the line in the figure on the right. The reason is that the vertical distances are, on the whole, smaller. The line that fits best is the line for which the sum of squared vertical distances is as small as possible. This line is called the least-squares regression line.





Equation of the Least-Squares Regression Line

Given ordered pairs (x, y), with sample means $\bar{x}$ and $\bar{y}$, sample standard deviations s_x and s_y, and correlation coefficient r,
the equation of the least-squares regression line for predicting y from x is
$\hat{y} = b_0 + b_1 x$

where $b_1 = r\frac{s_y}{s_x}$ is the slope and

$b_0 = \bar{y} - b_1\bar{x}$ is the y-intercept.

Vocabulary:

- The variable we want to predict (in this case, selling price) is called the **outcome variable**, or **response variable**.
- The variable we are given is called the **explanatory variable**, or **predictor variable**.
- In the equation of the least-squares regression line, x represents the explanatory variable and y represents the outcome variable.

Least Squares Regression Lines on the the TI-84 PLUS

Before computing the least-squares regression line, a one-time calculator setting should be modified to correctly configure the calculator to display the correlation coefficient. The following steps describe how to do this.

Step 1: Press **2nd, 0** to access the calculator catalog.

Step 2: Scroll down and select **DiagnosticOn**.

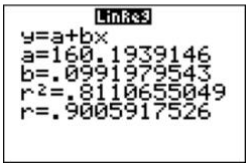
Step 3: Press **Enter** twice.

The following steps describe how to compute the least-squares regression line using the TI-84 PLUS calculator for the house size and selling price data.

Step 1: Enter the x-values in to **L1** and the y-values into **L2**.

Step 2: Press **STAT** and the right arrow key to access the **CALC** menu.

Step 3: Select the **LinReg(a+bx)** command. Verify that **L1** is entered in the **Xlist** field and **L2** in the **Ylist** field. Select **Calculate**.



The equation of the least-squares regression line is $\hat{y} = 160.1939 + 0.0992x$.

Objective 2: Use the Least-Squares Regression Lines to Make Predictions

Example 2
The equation of the least-squares regression line for predicting selling price from size is $\hat{y} = 160.1939 + 0.0992x$. Predict the selling price of a house of size 2800 sq. ft.

Objective 3: Interpret predicted values, the slope, and the y-intercept of the least-squares regression line

Interpreting the predicted value $\hat{y}$
The predicted value $\hat{y}$ can be used to estimate the average outcome for a given value of the explanatory variable.
(i.e.) For any given value of x, the value $\hat{y}$ is an estimate of the average y-value for all points with that x-value.

Example 3
Interpret the $\hat{y}$ value that you obtained in Example 2.

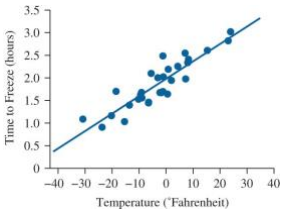
Interpreting the y-intercept b_0
The y-intercept b_0 is the point where the line crosses the y-axis. This has a practical interpretation only when the data contain both positive and negative values of x.

- If the data contain both positive and negative x-values, then the y-intercept is the estimated outcome when the value of the explanatory variable x is 0.
- If the x-values are all positive or all negative, then the y-intercept b_0 does not have a useful interpretation.

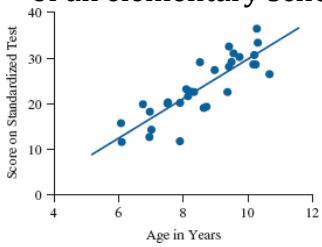
Example 4

For each of the following plots, interpret the y-intercept of the least-squares regression line, if possible. If not possible, explain why not.

a. The least-squares regression line is $\hat{y} = 1.98 + 0.039x$, where x is the temperature in a freezer in degrees Fahrenheit, and y is the time it takes to freeze a certain amount of water into ice.



b. The least-squares regression line is $\hat{y} = -13.586 + 4.340x$, where x represents the age of an elementary school student and y represents the score on a standardized test.

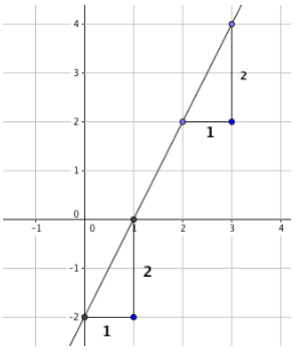


Interpreting the slope b_1

We know

$$Slope = \frac{Rise}{Run}$$

Slope tells you the change in y for a <u>unit change</u> in x.



- If the values of the explanatory variable for two individuals differ by 1, their predicted values will differ by b_1.
- If the values of the explanatory variable differ by an amount d, then their predicted values will differ by $b_1 \cdot d$.

Example 5
Two houses differ in size by 150 square feet. By how much should we predict their prices to differ?

Example 6
At the final exam in a statistics class, the professor asks each student to indicate how many hours he or she studied for the exam. After grading the exam, the professor computes the least-squares regression line for predicting the final exam score from the number of hours studied.
The equation of the line is $\hat{y} = 50 + 5x$

(a) Amy studied for 6 hours. What do you predict her exam scores to be?

(b) Emma studied for 3 hours longer than Jeremy. How much higher do you predict Emma's score would be than Jeremy?

Important Note while interpreting the slope:
The least squares regression line does not predict the result of changing the explanatory variables.

Example 7
A study is done in which a sample of men were weighed, and then each man was tested to see how much weight he could lift.
Explanatory variable, x = Weight of a man
Outcome, y = Amount he could lift.
The least-squares regression line was found to be $\hat{y} = 50 + 0.6x$
Slope = 0.6
So, Joe decides, "If I gain 10 pounds of weight, I will be able to lift 6 pounds more."
Is he right?

What is the correct interpretation of slope?

Example 8

The following table presents interest rates, in percent, for 30-year and 15-year fixed rate mortgages, for January through December 2012.

30-Year	3.92	3.89	3.95	3.91	3.80	3.68	3.55	3.60	3.47	3.38	3.35	3.35
15-Year	3.20	3.16	3.20	3.14	3.03	2.95	2.85	2.86	2.78	2.69	2.66	2.66

(a) Compute the least-squares regression line for predicting the 15-year rate from the 30-year rate.

(b) Is it possible to interpret the y-intercept?

(c) If the 30-year rate differs by 0.3 percent from one month to the next, by how much would you predict the 15-year rate to differ?

(d) Predict the 15 year rate for a month when the 30-year rate is 3.5 percent?

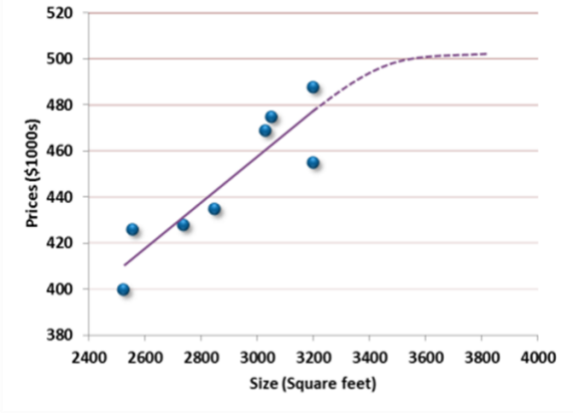
Sec 4.3 Features and Limitations of The Least-Squares Regression Line

Objectives

1. Understand the importance of avoiding extrapolation
2. Compute residuals and state the least-squares property
3. Construct and interpret residual plots
4. Determine whether outliers are influential
5. Compute and interpret the coefficient of determination

Objective 1: Understand the importance of avoiding extrapolation

Making predictions for values of the explanatory variable that are outside the range of the data is called **extrapolation.**
In general, it is best practice not to use the least-squares regression line to make predictions for x-values that are outside the range of the data because the linear relationship may not hold there.



Objective 2: Compute Residuals and State the Least-Squares Property

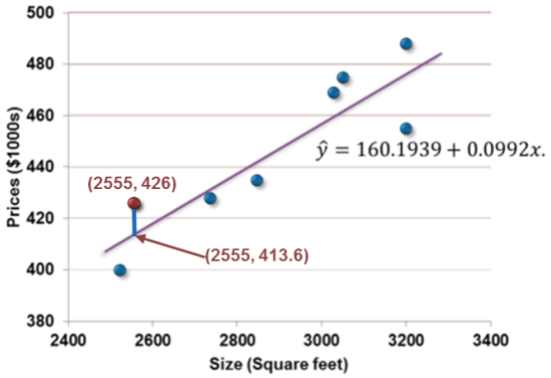
Given a point $(x,\ y)$ on a scatterplot, and the least-squares regression line

$$\hat{y} = b_0 + b_1 x$$

the residual for the point $(x,\ y)$ is the difference between the observed value y and the predicted value $\hat{y}$.

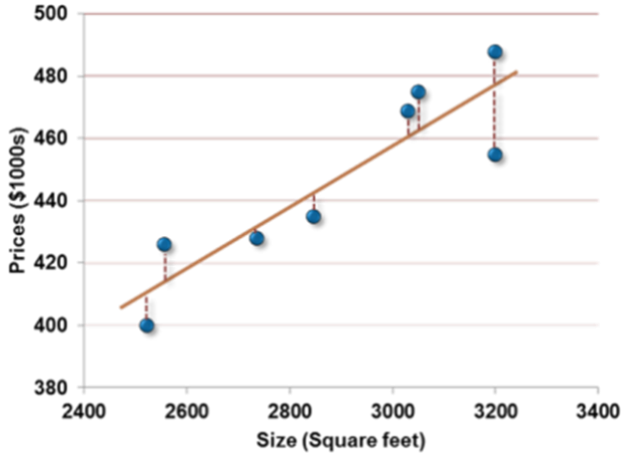
Residual = y - $\hat{y}$

For the least-squares regression line for predicting the selling price from house size $\hat{y} = 160.1939 + 0.0992x$, we may compute the residual for the point (2555, 426).
The predicted value $\hat{y}$ is $\hat{y} = 160.1939 + 0.0992(2555) = 413.6$.
The residual is $y - \hat{y} = 426 - 413.6 = 12.4$.



The Least-Squares Property

The magnitude of the residual is just the vertical distance from the point to the least-squares line.
The least-squares line is the line for which the sum of the squared vertical distances is as small as possible.
It follows that if we square each residual and add up the squares, the sum is less for the least-squares regression line than for any other line.
This is known as the **least-squares property**.



Objective 3: Construct and Interpret Residual Plots

Residual Plot

A **residual plot** is a plot in which the residuals are plotted against the values of the explanatory variable x.

- When two variables have a linear relationship, the residual plot will not exhibit any noticeable pattern.
- If the residual plot does exhibit a pattern, such as a curved pattern, then the variables do not have a linear relationship, and the least-squares regression line should not be used.

Important Note:
Do not rely on the correlation coefficient to determine whether two variables have a linear relationship. Even when the correlation is close to 1 or to −1, the relationship may not be linear.
To determine whether two variables have a linear relationship, construct a scatterplot or a residual plot.

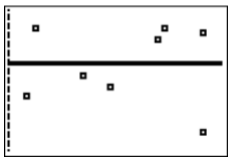
RESIDUAL PLOTS ON THE THE TI-84 PLUS

The following steps will create a residual plot for the house size and selling price data on the TI-84 PLUS.

Step 1: Enter the x-values into **L1** and the y-values into **L2**. Run the **LinReg(a+bx)** command.

Step 2: Press **2nd**, **Y=**, then **1** to access the Plot 1 menu. Select **On** and the scatterplot type. Enter the residuals for the **Ylist** field by pressing **2nd**, **STAT** and then **RESID**.

Step 3: Press **Zoom, 9** to view the plot.

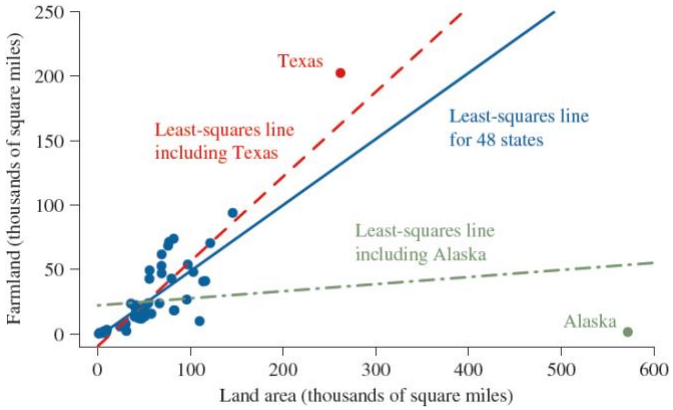


Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Objective 4: Determine whether Outliers are Influential

An **influential point** is a point that, when included in a scatterplot, strongly affects the position of the least-squares regression line.

Consider a scatterplot of farmland versus total land area for U.S. states.



Line L1 on the plot is the least-squares regression line computed for the 48 states not including Texas or Alaska.

Line L2 is the least-squares regression line for 49 states including Texas. Including Texas moves the line somewhat.

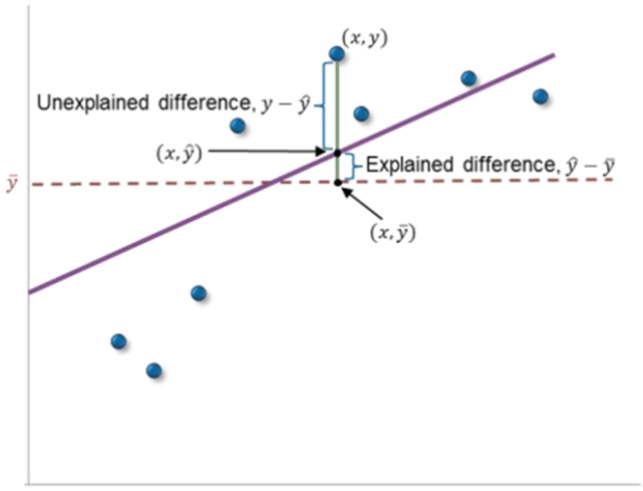
Line L3 is the least-squares regression line for 49 states including Alaska. Including Alaska causes a big shift in the position of the line.

Influential points are troublesome, because the least-squares regression line is supposed to summarize all the data, rather than reflect the position of a single point.
When a scatterplot contains outliers, the least-squares regression line should be computed both with and without each outlier, to determine which outliers are influential.
If there is an influential point, the least-squares regression line should be computed both with and without the point, and the equations of both lines should be reported

Objective 5: Compute and Interpret the Coefficient of Determination

Consider the least-squares line and the line $y = \bar{y}$.
For any point (x, y), $y - \bar{y}$ can be split into two parts.:

- The first part, $\hat{y} - \bar{y}$, the difference between the central value $\bar{y}$ and the predicted value $\hat{y}$, is called the **explained difference** and represents the difference explained by the least-squares line.
- The second part, $y - \hat{y}$, is the difference between the observed value y and the predicted value $\hat{y}$, which is just the residual. This difference is caused by factors unrelated to the least-squares regression line. It is called the **unexplained difference**.



The better the least-squares predictions are, the smaller the unexplained differences will be. We measure the size of the unexplained differences by squaring them and adding them together. This quantity is called the **unexplained variation**.

The **explained variation** is found similarly with the explained differences.

Coefficient of Determination

When two variables have a linear relationship, the correlation coefficient r tells how strong the relationship is.
The measure most often used to measure how well the least-squares regression line fits the data is r^2. The closer r^2 is to 1, the closer the predictions made by the least-squares regression line are to the actual values, on average.
The quantity r^2 is called the **coefficient of determination**.

Coefficient of Determination

$$r^2 = \frac{Explained\ Variation}{Total\ Variation}$$

Thus, r^2 measures the proportion of the total variation that is explained by the least-squares regression line.

Example
The correlation between size and selling price for the following data was computed to be $r = 0.9005918$. What is the coefficient of determination? How much of the variation in selling price is explained by the least-squares regression line?