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

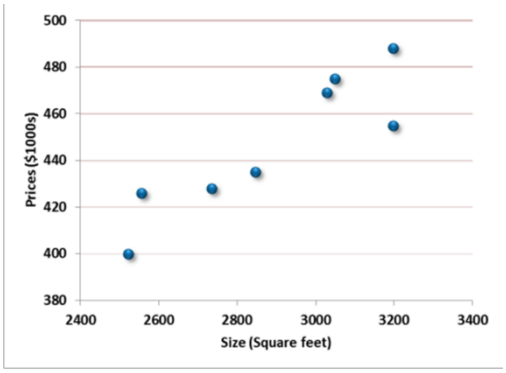
SCATTERPLOT

Suppose that a real estate agent wants to study the relationship between the size of a house and its selling price. It is reasonable to suspect that the selling price is related to the size of the house. Specifically, we expect that houses with larger sizes are more likely to have higher selling prices. A good way to visualize a relationship like this is with a **scatterplot**. In a scatterplot, each individual in the data set contributes an ordered pair of numbers, and each ordered pair is plotted on a set of axes.

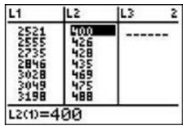
EXAMPLE: The following table presents the size in square feet and the selling price in thousands of dollars, for a sample of houses in a suburban Denver neighborhood. Construct a scatterplot for the data.

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

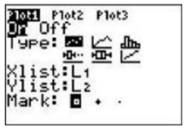
SOLUTION:



SCATTERPLOTS ON THE TI-84 PLUS



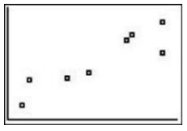
The following steps will create a scatterplot for the house sizes and prices data on the TI-84 PLUS.



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

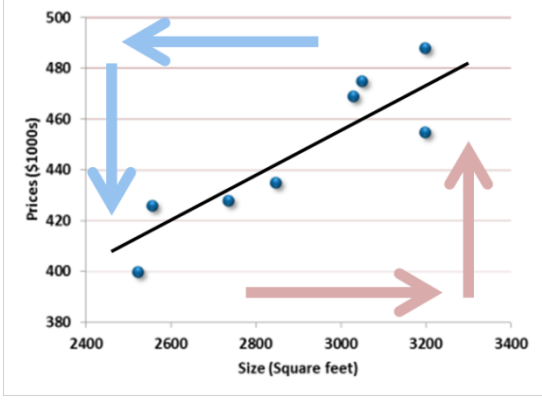
Step 2: Press **2nd,Y=**, then **1** to access the Plot1 menu. Select **On** and the scatterplot type.

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



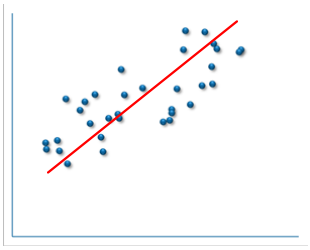
POSITIVE LINEAR ASSOCIATION

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.

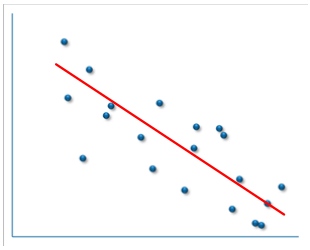


OTHER TYPES OF ASSOCIATION

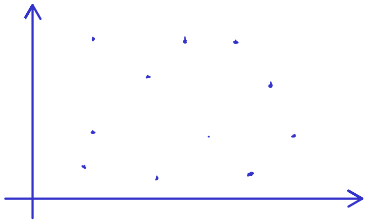
Two variables are **positively associated** if large values of one variable are associated with large values of the other.



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.



No linear association

OBJECTIVE 2
COMPUTE THE CORRELATION COEFFICIENT

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: Use the data in the following table to compute the correlation between size and selling price.

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

SOLUTION: L1: size L2: selling price

STAT CALC 4: LinReg(ax + b)

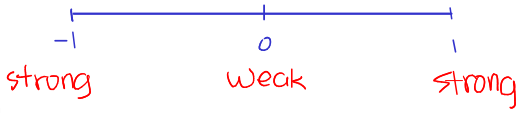
x list : L1
y list : L2
freq list : blank

correlation coefficient r:

r = 0.9005918

OBJECTIVE 3
INTERPRET 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. (There is no linear association)
- 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.

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.

YOU SHOULD KNOW …

- How to construct and interpret scatterplots
- The difference between positive, negative, linear, and nonlinear associations
- How to compute and interpret the correlation coefficient
- The difference between correlation and causation

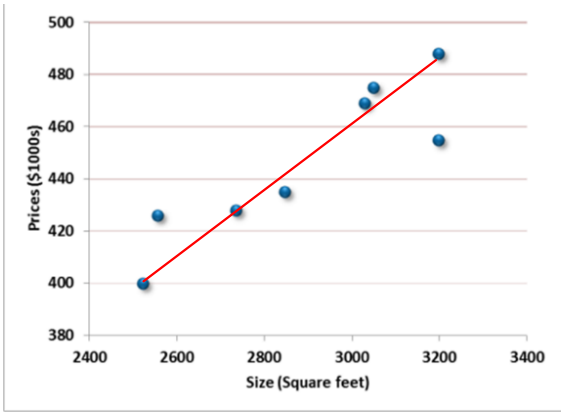
OBJECTIVES

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

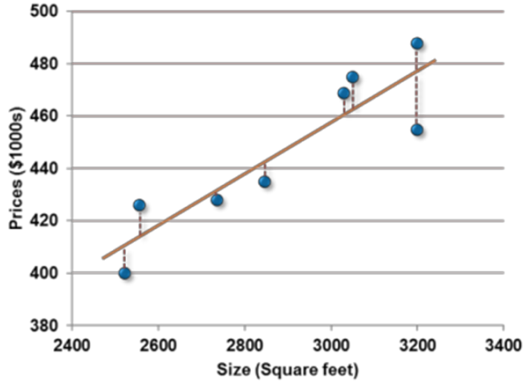
The table presents the size in square feet and selling price in thousands of dollars for a sample of houses. In the previous section, we concluded that there is a strong positive linear association between size and sales price. We can use these data to predict the selling price of a house based on its size.

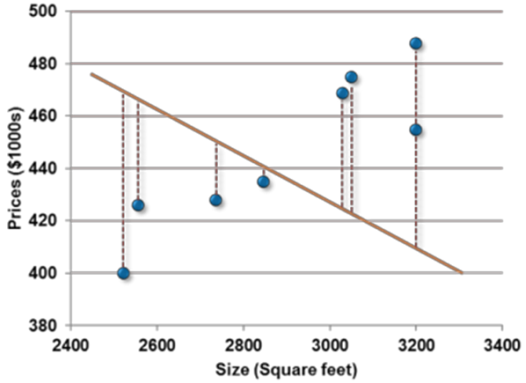
Size (Square Feet)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455



LEAST-SQUARES REGRESSION LINE

The figures present scatterplots of the previous 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. $\hat{y} = ax + b$

The variable we want to predict (in this case, selling price) is called the outcome variable, or response variable, and 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.

EXAMPLE: Compute the least-squares regression line for predicting selling price from size.

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

SOLUTION:

L1: size
L2: selling price

LinReg(ax+b) (L1, L2)
a = 0.099198
b = 160.1939

The equation of the least-squares regression line is

$\hat{y} = ax + b$

$\hat{y} = 0.099198x + 160.1939$

LSR LINES ON THE THE TI-84 PLUS

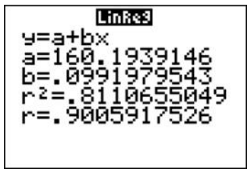
Least-squares regression lines are usually computed with technology rather than by hand. Before computing the least-squares regression line, a one-time calculator setting should be modified to correctly configure the calculator to display the correlation coefficient. The following steps describe how 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 LINE TO MAKE PREDICTIONS

PREDICTED VALUE

We can use the least-squares regression line to predict the value of the outcome variable by substituting a value for the explanatory variable in the equation of the least-squares regression line.

EXAMPLE: 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.

SOLUTION: To predict the selling price of a house size $x = 2800$ sq ft, we use $\hat{y} = 160.1939 + 0.0992x$

$= 160.1939 + 0.0992(2800)$

$= 437.9539 \approx 438$

The predicted selling price is 438 thousand dollar

OBJECTIVE 3
INTERPRET PREDICTED VALUES, THE SLOPE, AND THE y–INTERCEPT OF THE LEAST-SQUARES REGRESSION LINE

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

EXAMPLE: Use the equation of the least-squares regression line for predicting selling price from size $\hat{y} = 160.1939 + 0.0992$ to estimate the average price of all houses whose size is 3000 sq. feet.

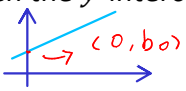
SOLUTION: $\hat{y} = 160.1939 + 0.0992(3000) = 457.8 \approx 458$
We estimate the average price of a house of 3000 sq feet to be 438 thousand dollars

INTERPRETING THE y-INTERCEPT b_0

$\hat{y} = b_1x + b_0$ or $\hat{y} = b_0 + b_1x$

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.

INTERPRETING THE SLOPE b_1

Recall the slope $m = \frac{y_2 - y_1}{x_2 - x_1} \Rightarrow y_2 - y_1 = m(x_2 - x_1)$ ($\Rightarrow$ implies; m = b_1)

If the x-values of two points on a line differ by 1, their y-values will differ by an amount equal to the slope of the line. This fact enables us to interpret the slope b_1 of the least-squares regression line. 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: Two houses differ in size by 150 square feet. By how much should we predict their prices to differ?

SOLUTION: The slope of the least-squares regression line is $b_1 = 0.0992$
We predict the prices to differ by $b_1 d = 0.0992(150) = 14.9$ thousand dollars

YOU SHOULD KNOW …

- The definitions of outcome or response variables and explanatory or predictor variables
- How to compute the least-squares regression line
- How to use the least-squares regression line to make predictions
- How to interpret the predicted value $\hat{y}$, the y-intercept b_0, and the slope b_1 of a least-squares regression line

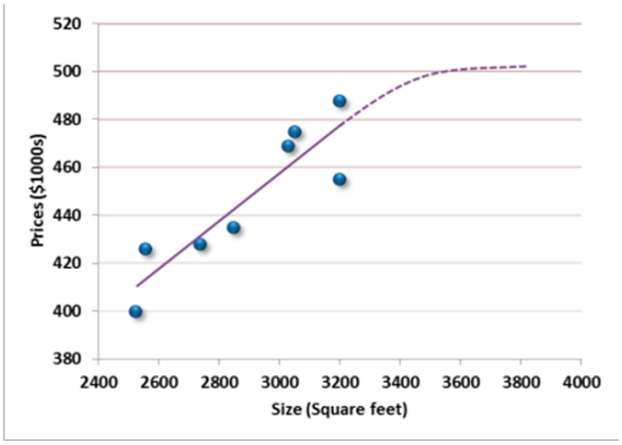
SECTION 4.3: FEATURES AND LIMITATIONS OF THE LEAST-SQUARES REGRESSION LINE

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.



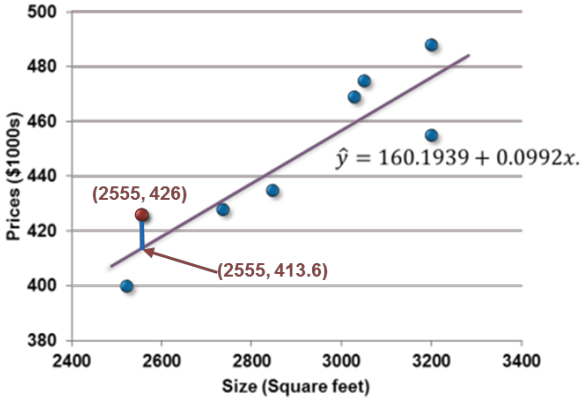
OBJECTIVE 2
COMPUTE RESIDUALS AND STATE THE LEAST-SQUARES PROPERTY

RESIDUALS

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

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

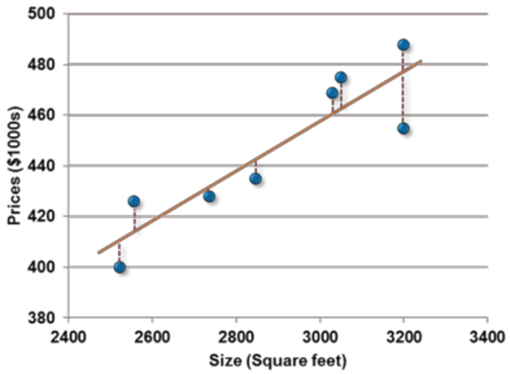
the residual for the point (x, y) is the difference between the observed value y and the predicted value $\hat{y}$.
For the least-squares regression line for predicting the selling price from house size $\hat{y} = 160.1939 + 0.0992x$, we may compute the residual for the point (2555, 426). The predicted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



SECTION 4.3: FEATURES AND LIMITATIONS OF THE LEAST-SQUARES REGRESSION LINE

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 PLOTS

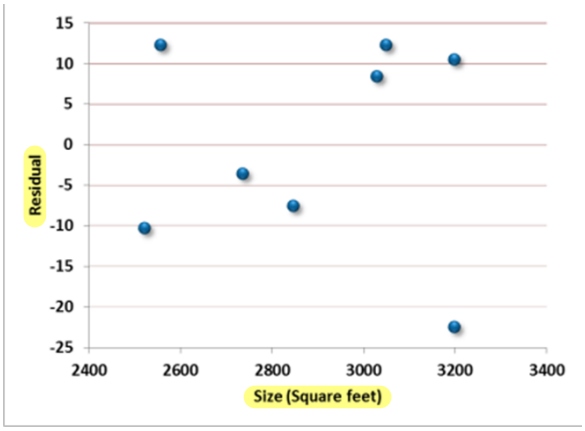
A **residual plot** is a plot in which the residuals are plotted against the values of the explanatory variable x.

- 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.

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.

EXAMPLE: The least-squares regression line for predicting selling price from house size is $\hat{y} = 160.1939 + 0.0992x$. The residuals and the residual plot are shown.

Size (Square Feet)	Selling Price y ($1000s)	Predicted Value $\hat{y}$	Residual $y - \hat{y}$
2521	400	410.27	–10.28
2555	426	413.65	12.35
2735	428	431.51	–3.51
2846	435	442.52	–7.52
3028	469	460.57	8.43
3049	475	462.66	12.35
3198	488	477.44	10.56
3198	455	477.44	–22.44

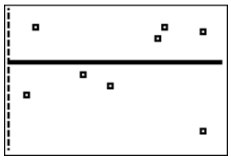


The residual plot exhibits no noticeable pattern, so use of the least-squares regression line is appropriate.

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.



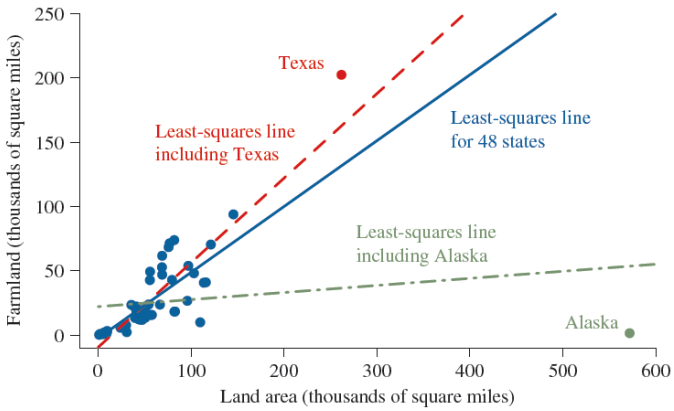
Step 2: Press **2nd**, **Y=**, then **1** to access the Plot 1 menu. Select **On** and the scatterplot type. Enter the residuals for the **Ylist** field by pressing **2nd**, **STAT** and then **RESID**.

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

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



The red dotted line is the least-squares regression line for 49 states including Texas. Including Texas moves the line somewhat. The green dash-dot line is the least-squares regression line for 49 states including Alaska. Including Alaska causes a big shift in the position of the line.

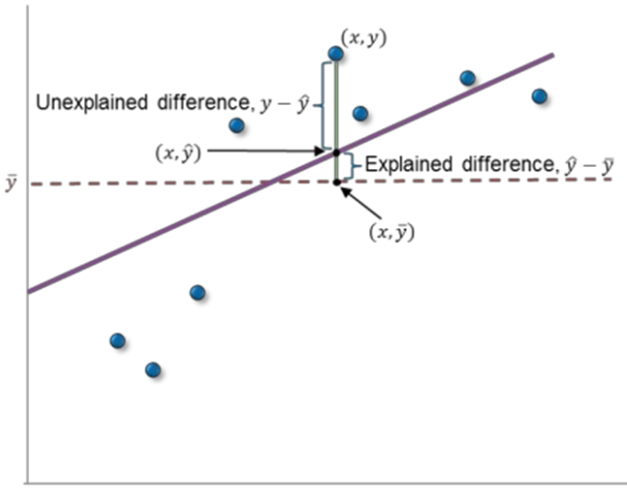
Influential points are troublesome, because the least-squares regression line is supposed to summarize all the data, rather than reflect the position of a single point.

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 $y = \bar{y}$. ← horizontal line 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:

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

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?

SOLUTION: The coefficient of determination is

$r^2 = (0.9005918)^2 = 0.811$

Interpretation: ("how much variation")

Therefore, 81.1% of variation in selling price is explained by the least-square regression line

YOU SHOULD KNOW …

- That extrapolation outside of the range of data should generally be avoided
- How to compute residuals
- The least-squares property
- How to construct and use a residual plot to determine whether it is appropriate to use least-squares regression
- What an influential point is and how they can affect least-squares regression lines
- How to compute and interpret the coefficient of determination