A correlation exists between two quantitative variables when there is a statistical relationship between them.

**Example 1: Golf Ball Height versus Distance:**
If you hit a golf ball and follow its position in the air as it travels, there is a correlation between its height above ground and the horizontal distance it has traveled. There IS A CORRELATION, but the relationship can NOT be modeled by a straight line.
There is NO "LINEAR CORRELATION".

**Example 2: Student College Fees versus Credits Enrolled:**
Community college fees are calculated based on a constant cost per credit unit. There is a CORRELATION between the number of credits a student enrolls in and the fees when attending community college.
This relationship follows a straight line – so there is a LINEAR CORRELATION between the number of credit units and the total course fees.
NOTE: The linear model does not extend past the data on the graph!

**Example 3: College Faculty Size versus Student Size:**
There is a LINEAR CORRELATION between the number of students and number of faculty for this sample of the 8 local community colleges in Santa Clara and Santa Cruz counties. The points lie near a line but not exactly on the line.
(Based on data for Fall, 2008 for Santa Clara and Santa Cruz county community colleges)

******** YOUR FIRST STEP SHOULD BE TO ALWAYS LOOK AT THE SCATTERPLOT TO VISUALLY SEE IF A LINE LOOKS LIKE A GOOD FIT TO THE DATA.

The points should be randomly scattered about the line. Obvious patterns that are not random often may suggest that some other model, perhaps a curve such as a parabola or exponential curve, would be a better fit to the data.
The process of finding the best fit line to fit given bivariate data \((x, y)\) is called linear regression.

- \(X\) represents the INDEPENDENT VARIABLE (input variable, horizontal variable)
- \(Y\) represents the DEPENDENT VARIABLE (output variable, vertical variable)

The best fit line \(\hat{y} = a + bx\) is called the least squares regression line or regression line or line of best fit where \(\hat{y} = a + bx\) is in slope-intercept form \((a\) is the vertical intercept; \(b\) is the slope).

What makes a line a best fit line? ----Least Squares Criteria for the Best Fit Line
The residual \(y - \hat{y}\) is the vertical “error” between the observed data value and the line.

Definition of Best Fit Line: The best fit line is the line for which \(\text{SSE} = \sum(y - \hat{y})^2\) is minimized.

SSE is the sum of the squares of the residuals, also called Sum of the Squared Errors.

The best fit criteria says to find the line that makes the SSE as small as possible

Any other line that you might try to fit through these points will have the sum of the squared residuals \(\text{SSE} = \sum(y - \hat{y})^2\) larger than the \(\text{SSE} = \sum(y - \hat{y})^2\) for the best fit line.

The best fit line is called the least squares regression line.

The formulas that are used to find the best fit line are developed from techniques in calculus.

We will use technology (calculator or computer) to find the equation of the best fit line.

**Example 4:** The following graphs show the same 6 data points with two different lines. One is the best fit line.

a. Complete the tables to compare the SSE's and choose the least squares regression line.

\[^\hat{y} = 1 + 2x\] \[^\hat{y} = -1 + 3x\]

<table>
<thead>
<tr>
<th>(x)</th>
<th>(y)</th>
<th>(\hat{y})</th>
<th>((y - \hat{y}))</th>
<th>((y - \hat{y})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>7</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Add up the \((y - \hat{y})^2\) column \(\text{SSE} = \sum(y - \hat{y})^2 = 6\)

<table>
<thead>
<tr>
<th>(x)</th>
<th>(y)</th>
<th>(\hat{y})</th>
<th>((y - \hat{y}))</th>
<th>((y - \hat{y})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Add up the \((y - \hat{y})^2\) column \(\text{SSE} = \sum(y - \hat{y})^2 = ____\)

b. Which line has the smallest SSE? ________________ c. Which line is the best fit? ________________
The regression line can be used to predict the output value, \( \hat{y} \) (y-hat), for a specific input value, \( x \).

- \( \hat{y} \) is an estimate or prediction of the actual data output value, \( y \).
- \( \hat{y} \) is the value we would expect to get for \( y \), on average, for a specific value of input, \( x \).
- Predicting \( \hat{y} \) is ONLY VALID for an input value, \( x \), that lies in the DOMAIN of the sample data used to calculate the regression line. (x-values between the minimum and maximum x-values, inclusive)

MEASURING THE STRENGTH OF A LINEAR RELATIONSHIP:

- **CORRELATION COEFFICIENT:**
  \( r \) is a numerical measure of the strength of the LINEAR relationship between 2 variables based on sample data.

\[
 r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}
\]

We will use technology (calculator or computer) to calculate the sample correlation coefficient, \( r \).

- \(-1 \leq r \leq 1\). The sign of \( r \) is the same as the sign of the slope of the regression line.
- The closer the linear correlation coefficient \( r \) is to 0, the weaker the linear correlation.
- The farther the linear correlation coefficient \( r \) is from 0, (closer to \(-1\) or \(1\)), the stronger the correlation.

<table>
<thead>
<tr>
<th>( r )</th>
<th>Negative ( r ) and negative slope</th>
<th>0</th>
<th>Positive ( r ) and positive slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1)</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(0)</td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Perfect linear correlation of sample data
Use: \( \hat{y} = a + bx \)
Slope is negative.

No linear correlation of sample data
Use: \( \hat{y} = \bar{y} \)
Slope = 0

Perfect linear correlation of sample data
Use: \( \hat{y} = a + bx \)
Slope is positive.

- **COEFFICIENT OF DETERMINATION:** \( r^2 \)
  - \( r^2 \) is the square of the correlation coefficient.
  - \( 0 \leq r^2 \leq 1 \); but \( r^2 \) is usually stated as a percent between 0% and 100%.
  - \( r^2 \) is the percent (or proportion) of the total variation in the y values that can be explained by the variation in the x values, using the best fit line.
  - So, \( 1 - r^2 \), restated as a percent, tells you the percent of variation in the y values that is not explained by the linear relationship between \( x \) and \( y \). This variation may be due to other factors, or may be random, or a combination of both reasons. This variation is seen in the graph as the scattering of points about the line.
  - The closer the coefficient of determination, \( r^2 \), is to 1, the more reliable the regression line will be.
EXAMPLE 5: Scatterplots showing Positive and Negative Linear Correlation

interpreting Correlation Coefficient \( r \) and Coefficient of Determination \( r^2 \)

\[ r = 0.962 \]: There is a strong positive linear correlation between the number of students enrolled and number of faculty at a community college.

\( r \) and the slope are BOTH positive.
The number of faculty increases as enrollment increases.

\[ r^2 = 0.925 \]: 92.5% of the variation in faculty size is explained by variation in enrollment, using the best fit regression line.

\[ r = -0.696 \]: There is a moderate negative linear correlation between the price of this DVD and the number of DVDs sold.

\( r \) and the slope are BOTH negative.
The number of DVDs sold decreases as the price increases.

\[ r^2 = 0.485 \]: 48.5% of the variation in weekly sales of this DVD is explained by the variation in price, using the best fit regression line.

CHECKLIST: 10 SKILLS AND CONCEPTS YOU NEED TO LEARN IN THIS CHAPTER

1. Identify which variable is independent and which variable is dependent, from the context of the problem.
2. Know your calculator skills for items 3, 4, 5, 6, 9 below.
3. Create and use a scatterplot to visually determine if it seems reasonable to use a straight line to model a relationship between the two variables.
4. Find, interpret, and use the correlation coefficient to determine if a significant linear relationship exists and to assess the strength of the linear relationship.
5. Find and interpret the coefficient of determination to determine what portion of the variation in the dependent variable is explained by the variation in the independent variable, using the best fit line, and what portion of the variation in the dependent variable is not explained by the line.
6. Find and use the least squares regression line to model and explore the relationship between the variables, finding predicted values, finding residuals, analyzing relationship between the observed and predicted values.
7. Know when it is and is not appropriate to use the least squares regression line for prediction - The scattergram of data must be well modeled with a line, \( p \)-value < \( \alpha \), the value of \( x \) for which we want to predict an dependent value must be in the domain of the data used to construct the best fit line.
8. Write the slope as a ratio of both numbers and word units. Write a verbal interpretation of the slope as marginal change in context of the problem (see guideline later in these notes).
9. Analyze the existence (or non-existence) of possible outliers and influential points, (identifying outliers by graphically determining values more than two standard deviations away from the graph of the line.)
10. Understand the concept of the least squares criteria for determining the best fit line.
EXAMPLE 6: Students' exam 3 grades and final exam grades - Linear Correlation versus Reliability:
The data below explore the relationship between students' exam scores on the third exam and on the final exam. (Data from Collaborative Statistics, R. Bloom, B. Illowsky & S. Dean Ch. 12)

A random sample of \( n=5 \) students was selected to examine the relationship between exam 3 and final exam grades. The sample size was increased with 6 more randomly selected students, for a larger sample of \( n=11 \) students.

<table>
<thead>
<tr>
<th>( n = 5 ) students (( r = 0.662 ))</th>
<th>( n = 11 ) students (( r = 0.663 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>x (third exam score)</td>
<td>y (final exam score)</td>
</tr>
<tr>
<td>65 66 75 70 69</td>
<td>175 123 198 163 159</td>
</tr>
</tbody>
</table>

Both samples have almost the same correlation coefficient \( r \).

As a result of the increased sample size, the data from the larger sample appears more reliable with a larger coefficient of determination, \( r^2 \).

Generally, removing or changing a single data value from the smaller sample will change the position of the linear regression line more than if the same point were removed or changed for a larger sample.

The reliability of the linear relationship depends on the number data points, \( n \), in the sample as well as the value of \( r \).

Testing The Significance Of The Sample Correlation Coefficient:

- Enables us to decide whether the linear relationship in the sample data is strong enough and reliable enough to use as an estimate of the model for a linear relationship for the whole population.

- \( \rho = \text{population correlation coefficient} \) (lower-case Greek letter "rho")
  - \( \rho \) is the population parameter. (unknown, calculated from all population data)

- \( r = \text{sample correlation coefficient} \)
  - \( r \) is the sample statistic. (known; calculated from sample data)
  - \( r \) is the best point estimate of \( \rho \).

- The hypothesis test lets us make a decision about the value of the population correlation coefficient, \( \rho \), based on sample data.

- We will decide:
  - if \( \rho \) is "significantly different from 0" OR
  - if \( \rho \) is "not significantly different from 0" ("close to 0").

Two Methods: p-value approach (done in class, in textbook and in chapter notes)
- critical value approach (in textbook and in chapter notes – not done in class)
**HYPOTHESIS TEST** Approach to Test Significance Of Sample Correlation Coefficient $r$

**Hypotheses:**
- $H_0: \rho = 0$  *(there is NOT a linear relationship between $x$ and $y$ in the population)*
- $H_a: \rho \neq 0$  *(there is a linear relationship between $x$ and $y$ in the population)*

The p-value tells us how likely a given sample correlation coefficient, $r$, will occur assuming the population parameter $\rho = 0$, in other words, assuming no linear relationship in the population.

The **smaller** the p-value, then
  - the more significant the linear relationship in the population,
  - the greater the difference between the statistic, $r$, and the assumed parameter $\rho=0$,
  - the closer to -1 or 1 the sample linear correlation coefficient, $r$ will be.
  - the further from 0 the sample linear correlation coefficient, $r$ will be.

We need to decide if the sample correlation coefficient for our set of sample data is far enough away from 0 for us to decide that the population correlation coefficient is unlikely to be 0.

**If the p-value is small, then the sample correlation coefficient $r$ is “far enough away from 0 to:”**

- **DECIDE** to Reject the null hypothesis, $H_0: \rho = 0$.
- **CONCLUDE** the data show strong enough evidence to believe $H_a: \rho \neq 0$,
  - the sample correlation coefficient $r$ is significant,
  - the population correlation coefficient $\rho$ is not equal to 0,
  - we can use the linear equation $\hat{y} = a + bx$ to estimate (predict) $y$ based on a given $x$ value.

➤ This means that the linear relationship in the sample data is strong and reliable enough to indicate that the linear relationship is likely to be true of the population. If the scatterplot also shows that a line is a reasonably appropriate model of this data, then we can model the relationship using the best fit regression line obtained from these data.

**If the p-value is larger than the significance level alpha, then the sample correlation coefficient $r$ is NOT sufficiently “far from 0”:**

- **DECIDE** to Not Reject the null hypothesis, $H_0: \rho = 0$.
- **CONCLUDE** the data do not show strong enough evidence to believe $H_a: \rho \neq 0$,
  - the sample correlation coefficient $r$ is not significant,
  - the population correlation coefficient $\rho$ is equal to 0,
  - and without any further investigation to seek better nonlinear models, we can only use $\hat{y} = \bar{y}$, to estimate $y$ based on any given $x$ value. (Remember, $\bar{y}$ is the average of all $y$ values.)

➤ This means that the linear relationship in the sample data is NOT strong and reliable enough to indicate that the linear relationship is likely to be true of the population. We can NOT model the relationship using the best fit regression line obtained from these data.
HYPOTHESIS TEST Approach to Test Significance Of Sample Correlation Coefficient $r$

**EXAMPLE 6 (Cont.):** Relationship between students’ exam 3 grades and final exam grades.

$\rho$ = population correlation coefficient

$r$ = sample correlation coefficient

Hypotheses:  
Ho : $\rho = 0$  (there is NOT a linear relationship between $x$ and $y$ in the population)  
Ha : $\rho \neq 0$  (there is a linear relationship between $x$ and $y$ in the population)

<table>
<thead>
<tr>
<th>n = 5 pairs of data, $r = 0.6615$</th>
<th>n = 11 pairs of data, $r = 0.663$</th>
</tr>
</thead>
</table>
| $y = a + bx$  \ $\beta \neq 0 \text{ and } \rho \neq 0$  
$t = 1.527927$  
$p = 0.223986$  
df = 3  (df = n−2 = 5−2)  
a = −139.62$  
b = 4.403  
s = 22.69  
r$^2$ = .4376  
r = .6615  
| $y = a + bx$  \ $\beta \neq 0 \text{ and } \rho \neq 0$  
$t = 2.65756$  
p = 0.02615  
df = 9  (df = n−2 = 11−2)  
a = −173.513$  
b = 4.827  
s = 16.41  
r$^2$ = .43969  
r = .66309  |

Decision: Do not reject Ho  
Reason for Decision:  
p-value > $\alpha$  
0.224 > 0.05

Conclusion: $r$ is NOT SIGNIFICANT.  
We should NOT use the regression line obtained from the data.

Explanation of conclusion:  
We can’t conclude that $\rho \neq 0$. We believe that it may be true that $p=0$ for the population.  
($r$ is not significantly different from 0.)

Based on $n = 5$ data points and $r = 0.6615$, the linear relationship seen in the sample is not strong enough and/or reliable enough to conclude that the linear relationship is also true in the population.

|----------------------------------|----------------------------------|
| Decision:  Reject Ho  
Reason for Decision:  
p-value < $\alpha$  
0.026 < 0.05

Conclusion: $r$ is SIGNIFICANT.  
We must also look at the scatter plot to be sure that it shows a reasonable linear trend.  
We CAN USE the regression line from the data.

Explanation of conclusion:  
We conclude that $\rho \neq 0$ for the population.  
($r$ is significantly different from 0.)

Because $r$ is significant and the scatter plot shows a reasonable linear trend, for $n = 11$ data points and $r = 0.663$, the sample data show evidence that linear relationship is both strong and reliable enough to conclude that the linear relationship exists in the population. The regression line for this sample data of 11 students CAN be used to predict final exam scores.

Note: The p-value method is equivalent to the critical value method.

CV is the value of $t$ for which the area in both tails totals the significance level 0.05. Using this value of $t$ in the test statistic formula and algebraically solving for “$r = \_” gives the values contained in the critical value table.

What your calculator does for you:

Test statistic is $t = \frac{r\sqrt{n−2}}{\sqrt{1−r^2}}$

degrees of freedom is $df = n−2$

The p-value is 2-tailed probability for the distribution $t_{n−2}$ by using tcdf to find the area further out in both tails than the $\pm$ calculated values of the test statistic.
CRITICAL VALUE Approach To Test Significance Of Sample Linear Correlation Coefficient r - Algebraically compare correlation coefficient, r to the CRITICAL VALUE:

95% Critical Values (CV) of the Sample Correlation Coefficient, r

<table>
<thead>
<tr>
<th>Degrees of Freedom (n-2)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Values (+ and -)</td>
<td>.997</td>
<td>.950</td>
<td>.878</td>
<td>.811</td>
<td>.754</td>
<td>.707</td>
<td>.660</td>
<td>.632</td>
<td>.602</td>
<td>.576</td>
<td>.555</td>
<td>.532</td>
<td>.514</td>
<td>.497</td>
<td>.482</td>
</tr>
</tbody>
</table>

1. Determine the sample size, n.
2. Calculate the degrees of freedom, n−2.
3. Compare the sample correlation coefficient to the corresponding critical value.
   - If r is closer to a perfect linear fit (−1 or 1) than its corresponding critical value
     \((-1 < r \leq -CV < 0 \) or \(0 < CV \leq r < 1)\) then,
     → Decision: Reject Ho: \(\rho = 0\).
     → Conclusion: r is SIGNIFICANT and BELIEVE \(\text{Ha: } \rho \neq 0\).
     → We CAN USE the regression line from this data. \(\hat{y} = a + bx\)
   - If r is not closer to a perfect linear fit (−1 or 1) than its corresponding critical value,
     \((-1 < CV \leq r \leq 01 \) or \(0 \leq r \leq CV < 1)\) then,
     → Decision: Do Not Reject Ho: \(\rho = 0\).
     → Conclusion: r is NOT SIGNIFICANT and DO NOT BELIEVE \(\text{Ha: } \rho \neq 0\).
     → We CAN NOT USE the regression line from this data.

Instead, we must estimate with \(\hat{y} = \bar{y}\), a horizontal line.

EXAMPLE 6 (Cont.): Relationship between students' exam 3 grades and final exam grades.

Hypotheses:   Ho  : \(\rho = 0\) \((there is NOT a linear relationship between x and y in the population)\)
               Ha  : \(\rho \neq 0\) \((there is a linear relationship between x and y in the population)\)

<table>
<thead>
<tr>
<th>n = 5 pairs of data , (r = .6615)</th>
<th>n = 11 pairs of data , (r = .6631)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 5,</td>
<td>n = 11,</td>
</tr>
<tr>
<td>so degrees of freedom = 5 - 2 = 3</td>
<td>so degrees of freedom = 11 - 2 = 9</td>
</tr>
<tr>
<td>The critical values are ± 0.878.</td>
<td>The critical values are ± 0.602.</td>
</tr>
<tr>
<td>(r = 0.6615), a positive value making the corresponding critical value, CV = 0.878.</td>
<td>(r = 0.6631), a positive value making the corresponding critical value, CV = 0.602.</td>
</tr>
<tr>
<td>Graphically, we can show:</td>
<td>Graphically, we can show:</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\text{CV} = .878) (\text{CV} = .602)</td>
</tr>
</tbody>
</table>

Decision: Do not reject Ho

Reason for Decision:
0 < \(r = 0.6615\) < CV = 0.878 < 1

Conclusion: r is NOT SIGNIFICANT, \(\rho = 0\)

We should NOT use the regression line obtained from the data: \(\hat{y} = \bar{y}\)

Decision: Reject Ho

Reason for Decision:
0 < CV = 0.602 < \(r = 0.6631\) < 1

Conclusion: r is SIGNIFICANT, \(\rho \neq 0\)

We must also look at the scatter plot to be sure that it shows a reasonable linear trend.

We CAN USE the regression line from the data: \(\hat{y} = a + bx\)
CALCULATOR INSTRUCTIONS FOR TI-83, 83+, 84+:

DRAWING A SCATTERPLOT

TI-83, 83+, 84:
2nd STATPLOT 1

On Off
Type Highlight the scatterplot icon and press enter
Xlist: list with x variable
Ylist: list with y variable
Mark: select the mark you would like to use for the data points
ZOOM 9:ZoomStat

Use TRACE and the right and left cursor arrow keys to jump between data points and show their (x,y) values.

LINEAR REGRESSION T TEST

TI-83, 83+, 84+: STAT → TESTS → LinRegTTest

Xlist: enter list containing x variable data
Ylist: enter list containing y variable data
Freq: 1
β & ρ: ≠ 0 <0 >0 Highlight ≠ 0 ENTER
RegEq: Leave RegEq blank
Calculate Highlight Calculate; then press ENTER

Graph the best fit line on scatter plot with the equation found with the LinRegTTest:

Find equation of line \( \hat{y} = a + bx \)
using the values of a and b given on LinRegTTest calculator output.

TI-83, 83+, 84+:
Press Y=

Type the equation for a + bX into Y1.
(use X t θ n key to enter the letter X).

OUTPUT SCREEN

LinRegTTest
\( y = a + bx \)
\( \beta \neq 0 \text{ and } \rho \neq 0 \)
\( t = \text{test statistic} \)
\( p = p\text{-value} \)
\( df = n - 2 \)
\( a = \text{value of y-intercept} \)
\( b = \text{value of slope} \)
\( s = \text{standard deviation of residuals} \) \((y-\hat{y})\)
\( r^2 = \text{coefficient of determination} \)
\( r = \text{correlation coefficient} \)

TO IDENTIFY OUTLIERS
(Note: your text uses the term "potential outliers"): Graph 3 lines \( Y1 = a+ bx, Y2 = a+bx-2s, Y3 = a+bx+2s \)
on the same screen as the scatterplot.

Any data points that are above the top line or below the bottom line are OUTLIERS.

Data points that are between the lines are not outliers.

Use TRACE and the right and left arrow cursor keys to jump to the outliers to identify their coordinates.

The calculator's screen resolution may make it hard to tell if a point is inside or outside the lines if it is very close to the line or appears to be exactly on the line. If the graph does not give clear information, do the calculation numerically to determine if it is outside or inside the lines.
OUTLIERS IN LINEAR REGRESSION

An outlier is a data point that is unusually far away from the regression line.

An influential point is a data point that is far away from the other data points and strongly influences the best fit line.

Rough Rule of Thumb for Outliers: If a data point is more than two standard deviations away from the regression line, the data point may be considered an outlier. The standard deviation (s) used is the standard deviation of the residuals, or errors: \( (y – \hat{y}) \) (vertical distances between the data points and the line)

Outliers should be examined to see if they are correct and/or belong in the data set; then a decision can be made whether to leave the outlier in the data or remove the outlier from the data.

**EXAMPLE:**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>

Find the equation of the best fit line is \( \hat{y} = \)_________

We need to find and draw the lines that are two standard deviations away from (above and below) the regression line.

The standard deviation is \( s = \)_______

Find the outlier: The point __________ is an outlier.

This package weighed _______ pounds and had a shipping cost of $______.

**Graph the data and these lines on your calculator and draw the lines onto the graph here:**

\[
\begin{align*}
Y_1 &= .99 + .51X \\
Y_2 &= .99 + .51X - 2\times1.97 \\
Y_3 &= .99 + .51X + 2\times1.97
\end{align*}
\]

For outlier point (9,12), predict \( \hat{y} : \hat{y} = \)_________

For this 9 pound package the actual shipping cost is $____ and the line predicts a shipping cost of $______.

Does the observed data point lie above or below the best fit line?:

observed data point is _________ the best fit line

Does the line overestimate or underestimate the observed data value?

predicted value on the line _________ the observed data value

Find the residual: \( y – \hat{y} = \)______ – ________ = __________

**INTERPRETATION OF SLOPE:** slope =.51

For every 1 pound increase in weight, the shipping cost increases by $.51

OR The shipping cost increases at the rate of $.51 per additional pound of weight.

**INTERPRETATION OF COEFFICIENT OF DETERMINATION:**

\( r^2 = .555: \) 55.5\% of the variation in shipping costs of packages is explained by the variation in weight

\( 1 - r^2 = 1 -.555 = .445 \)

44.5\% of the variation in shipping costs of packages is NOT explained by variation in weight, and therefore is due to other factors and/or random variation
**GUIDELINE: INTERPRETATION OF THE SLOPE OF THE REGRESSION LINE**

Slope = \( b = \frac{\text{rise}}{\text{run}} = \frac{\text{change in } \hat{y}}{\text{change in } x} = \frac{\text{b unit change in } \hat{y}}{1 \text{ unit change in } x} \)

The interpretation of the slope must state the amount of CHANGE (and indicate if it is an increase or decrease) in the dependent variable \( y \) for a one unit increase in the independent variable \( x \).

It must be written to clearly explain this change in the words and appropriate units relating to the context of the problem, rather than just referring to "\( x \)" and "\( y \)."

**Interpretation of slope for examples used in chapter notes:**

Example 7:
Weights and shipping costs of packages: \( \text{slope} = 0.51 \)

- **Interpretation:** For every 1 pound increase in weight, the shipping cost increases by $0.51, on average.
- **OR**
- **Interpretation:** The shipping cost increases by $0.51 per additional pound of weight, on average.

Example 6 with \( n = 11 \) students:
Third Midterm Grade and Final Exam Grade: \( \text{slope} = 4.827 \)

- **Interpretation:** For every one additional point on the third midterm grade, we predict that the final exam grade increases by 4.827 points, on average.

Skills Practice Example 1: Community college student body and faculty size: \( \text{slope} = 0.032 \)

- **Interpretation:** The number of faculty increases by an average of 0.032 faculty for each additional student.
  
  *(In this case: The number of faculty increases by 0.032 faculty for each additional 100 students, on average)*

Skills Practice Example 2: Price of DVDs and sales of DVDs: \( \text{slope} = -18 \)

- **Interpretation:** The number of DVDs sold decreases by 18 DVDs for every $1 increase in price, on average.
- **OR**
- **Interpretation:** A $1 price increase reduces sales by 18 DVDs, on average.

Additional example of interpreting the slope of the regression line

**Math SAT scores and hours of studying:**
A study was conducted involving 20 students as they prepared for and took the Math section of the SAT Examination.

- **X:** the number of hours studied by one student for the Math SAT exam
- **Y:** the student's Math SAT score

\[ \hat{y} = 25.3x + 353.2 \]

- **Slope:** 25.3 Math SAT points per hour of studying = \( \frac{25.3 \text{ Math SAT points}}{1 \text{ hour of studying for the Math SAT}} \)

- **Interpretation:**
For each 1 hour increase in a student's study time for the Math SAT, we predict an increase of 25.3 points, on average, in the student's Math SAT score.

- **OR**

Studying an additional hour for the Math SAT exam is predicted to increase a student's score by 25.3 points, on average.

- **III** It is **INCORRECT** to just write: A student scores 25.3 points per hour of study for the Math SAT. *The reason this is incorrect is that it does not explicitly refer to change, increase or decrease.*
SKILLS PRACTICE: PROBLEM 1:

<table>
<thead>
<tr>
<th></th>
<th>X = Number of Students</th>
<th>Y = Number of Faculty</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Anza</td>
<td>26173</td>
<td>846</td>
</tr>
<tr>
<td>Foothill</td>
<td>20919</td>
<td>618</td>
</tr>
<tr>
<td>West Valley</td>
<td>13800</td>
<td>433</td>
</tr>
<tr>
<td>Mission</td>
<td>12814</td>
<td>411</td>
</tr>
<tr>
<td>San Jose City</td>
<td>11513</td>
<td>436</td>
</tr>
<tr>
<td>Evergreen</td>
<td>10936</td>
<td>330</td>
</tr>
<tr>
<td>Gavilan</td>
<td>9092</td>
<td>234</td>
</tr>
<tr>
<td>Cabrillo</td>
<td>16369</td>
<td>618</td>
</tr>
</tbody>
</table>

The data at left show the number of students enrolled and number of faculty at community colleges in Santa Clara County and Santa Cruz County. This data is from the state's community college website data bank for fall 2008.

1. Find the best fit line and write the equation of the line.
2. Graph a scatterplot of the data, showing the best fit line.
3. Find the correlation coefficient and the coefficient of determination.
4. Considering this data as a sample of all bay area community colleges, test the significance of the correlation coefficient. Show your work and clearly state your conclusion.

5. Write the interpretation of the coefficient of determination in the context of the data.
6. Write the interpretation of the slope of the regression line, in the context of the data.
7. How many faculty would be predicted at a college with 15000 students.
8. How many faculty are predicted for a college with 11,513 students?
   What is the residual (y – ŷ: difference between the observed y and predicted ŷ ) when x = 11,513?
   Did value predicted by the line overestimate or underestimate the observed value?

9. How many faculty are predicted for a college with 20,919 students?
   What is the residual (y – ŷ: difference between the observed y and predicted ŷ ) when x = 20,919?
   Did value predicted by the line overestimate or underestimate the observed value?

10. Would it be appropriate to use the line to predict the number of faculty at a community college with:
    4000 students? _________   14000 students? _________   40000 students? _______
    Explain why or why not.
SKILLS PRACTICE: PROBLEM 2

Do sales of a DVD increase when its price is lower?

The Videomart chain stores are selling a particular DVD at all their stores in the state. The price varies during different weeks, some weeks at full price, other weeks at discounted prices.

The manager recorded the price and sales for this particular DVD for a sample of 12 weeks.

(Sales have been rounded to the nearest 10; the price is the same at all store branches during the same week.)

\[ x = \text{price of this DVD during a one week period} \]
\[ y = \text{number of this DVD sold during a one week period} \]

<table>
<thead>
<tr>
<th>x ($)</th>
<th>y (DVDs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>370</td>
</tr>
<tr>
<td>15</td>
<td>400</td>
</tr>
<tr>
<td>15</td>
<td>330</td>
</tr>
<tr>
<td>16</td>
<td>380</td>
</tr>
<tr>
<td>18</td>
<td>250</td>
</tr>
<tr>
<td>13</td>
<td>340</td>
</tr>
<tr>
<td>16</td>
<td>350</td>
</tr>
<tr>
<td>15</td>
<td>310</td>
</tr>
<tr>
<td>16</td>
<td>360</td>
</tr>
<tr>
<td>18</td>
<td>260</td>
</tr>
<tr>
<td>13</td>
<td>380</td>
</tr>
<tr>
<td>18</td>
<td>290</td>
</tr>
</tbody>
</table>

1. Find the best fit line and write the equation of the line.
2. Graph a scatterplot of the data, showing the best fit line
3. Find the correlation coefficient and the coefficient of determination
4. Test the significance of the correlation coefficient. Show your work and clearly state your conclusion
5. Write the interpretation of the coefficient of determination in the context of the data.
6. Write the interpretation of the slope of the regression line, in the context of the data.
7. Predict the sales if the price is $15. How does this compare to the observed sales when the price is $15?
8. The sales manager asks you to predict sales if he offers a special sale price of $10 for one week. What should you answer?