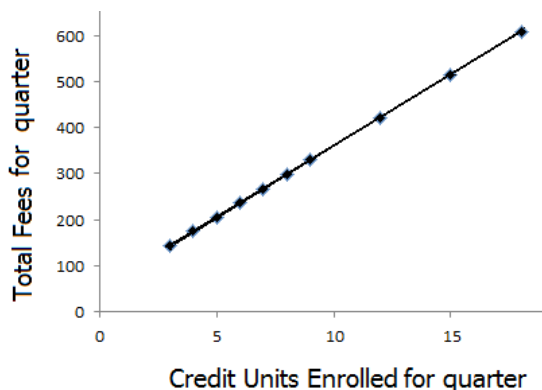# Chapter 12 : Linear Correlation and Linear Regression

Determining whether a linear relationship exists between two quantitative variables,
and modeling the relationship with a line, if the linear relationship is significant.

EXAMPLE 1.

At a community
college students pay a
basic fee of $50 per
quarter,
plus a fee of $31 per
credit unit:

| X = # of Credit Units	Y = Total Fees for quarter
3	143
4	174
5	205
6	236
7	267
8	298
9	329
12	422
15	515
18	608

EXAMPLE 2.

The data show the
relationship between
the number of
students and the
number of instructors
at a sample of 8 Bay
Area community
colleges during a
recent term.
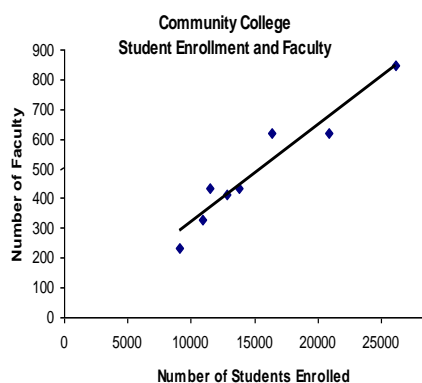
	X = Number of Students	Y = Number of Faculty
De Anza	26173	846
Foothill	20919	618
West Valley	13800	433
Mission	12814	411
San Jose City	11513	436
Evergreen	10936	330
Gavilan	9092	234
Cabrillo	16369	618

EXAMPLE 3.

A statistics instructor examined the relationship between her students' grades on a midterm exam and
their grade on the final exam, for a random sample of 11 students.
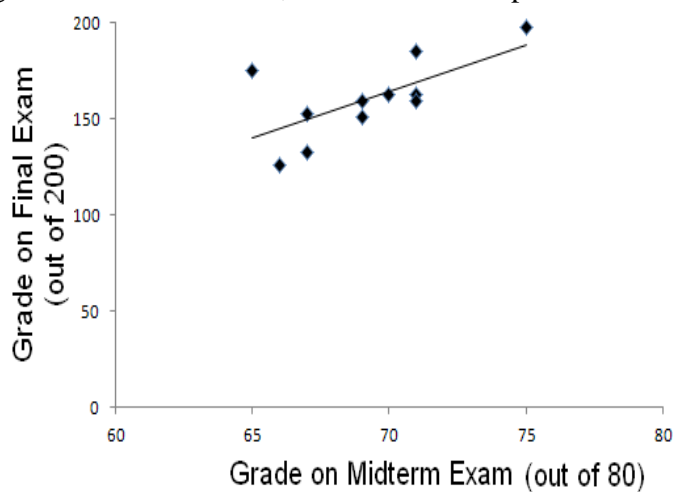
	X =Grade on Midterm Exam	Y = Grade on Final Exam
Student A	65	175
Student B	67	133
Student C	71	185
Student D	71	163
Student E	66	126
Student F	75	198
Student G	67	153
Student H	70	163
Student J	71	159
Student K	69	151
Student L	69	159

EXAMPLE 4.
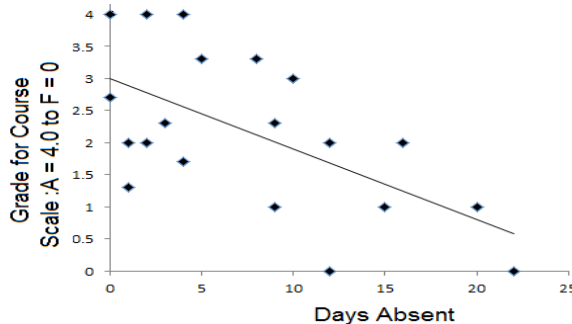An instructor examined the relationship between
X = number of absences a student has
    during a quarter
    (out of 54 classes for the quarter)
and
Y = student's grade for the course
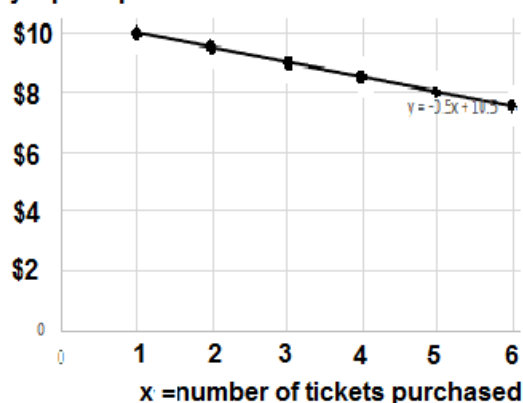   (scale of 0 to 4 where 4 = A and 0 = F)

Days Absent	0	0	1	1	2	2	3	4	4	5	8	9	9	10	12	12	15	16	20	22
Course Grade	3	4	2	1	4	2	2	2	4	3	3	2	1	3	0	2	1	2	1	0

EXAMPLE 5.
A sightseeing tour bus charges $10 per ticket.
Their "Family Tour" plan offers discounts per ticket that depend on the total number of tickets purchased, up to 6 tickets.


y = price per ticket

$y = -0.5x + 10.5$

x =number of tickets purchased

X = number of tickets in group	Y = price per ticket
1	$10.00
2	$9.50
3	$9.00
4	$8.50
5	$8.00
6	$7.50

EXAMPLE 6.
A golf ball is hit into the air from the ground. Its height above ground (y) and the horizontal distance (x) it has traveled are related by a parabolic curve.
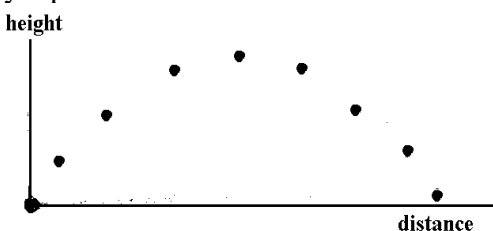
EXAMPLE 7.
Relationship between X = the age of a college student and Y = the number of cousins the student has:
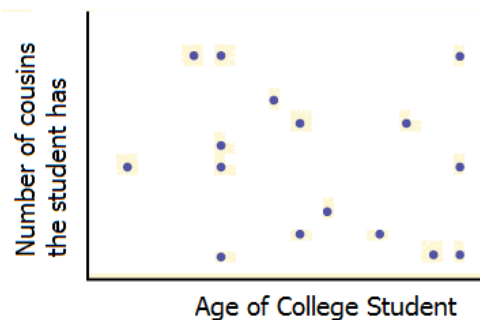
Before we can use the best fit line for a data set, we need to determine if a line is a good fit for the data.

SCATTER PLOT
- Create a scatterplot of the data using STATPLOT in your calculator
- Examine the scatterplot to see if a line appears to be a good model for the trend of the data.
  - ○ Is a line a reasonable model
  - ○ Might a curve be a better fit
  - ○ Does there appear to be no relationship at all between x and y.

DETERMINING IF A LINE IS A GOOD FIT TO THE DATA

CORRELATION COEFFICIENT r:
A number that measures the strength of the linear relationship between two quantitative variables.
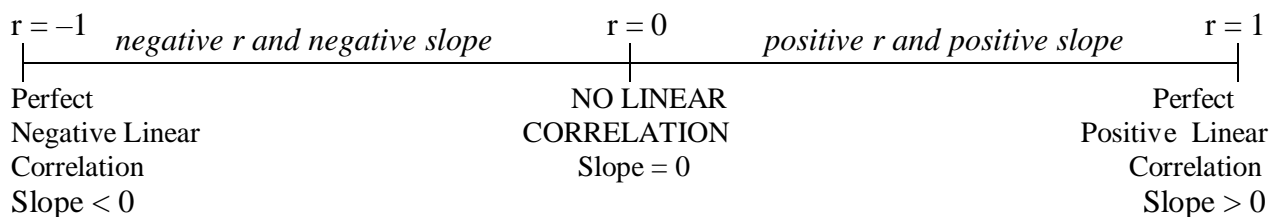The symbol for the correlation coefficient for a sample is r. $-1 \leq r \leq 1$

If the points all lie exactly on the line, the correlation coefficient is r = +1 or r = −1.
If there is no linear relationship between the variables, the correlation coefficient is r = 0 (you can **not**
fit a line to show the trend of the data points).
- The stronger the correlation and the more closely the points fit to the line, the closer r is to −1 or 1
  and the further r is from 0
- The weaker the correlation and the more scattered the points about the line, the closer r is to 0 .

The sign of r is the same as the sign of the slope of the best fit line
- If y increases as x increases, then the line slopes uphill and has a positive slope and r > 0
- If y decreases as x increases, then the line slopes downhill and has a negative slope and r < 0

r = −1   *negative r and negative slope*   r = 0   *positive r and positive slope*   r = 1

Perfect	NO LINEAR	Perfect
Negative Linear	CORRELATION	Positive Linear
Correlation	Slope = 0	Correlation
Slope < 0		Slope > 0

Formulas for Correlation Coefficient used by your calculator
Conceptually r examines the variation in x and y jointly compared to the variation in each variable separately

Theoretical formula defining r   "Easier" formula for doing calculations

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

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}\sqrt{n(\sum y^2)-(\sum y)^2}}$$

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

TRY-IT: **For Examples 1-7 on pages 1 & 2 the correlation coefficients are** (*in random order*) **:**
r = −1.0,   r = 0.66,   r = −0.61,   r = 0.96,   r = 1.0,   r = 0
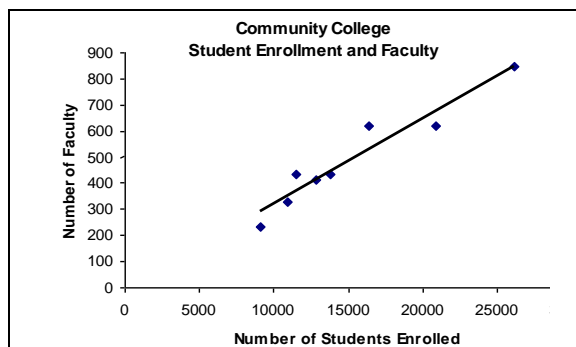**For each graph determine which of value of r above best corresponds to the**
graph and write the value of r on the graph in the form "r = ____".

COEFFICIENT OF DETERMINATION: r^2
- r^2 is the square of the correlation coefficient, so $0 \leq r^2 \leq 1$, but r^2 is usually stated as a percent
  between 0% and 100%
- The closer the coefficient of determination, r^2 is to 1, the more reliable the regression line will be
- **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.
- $1 - r^2$ is 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. This variation is
  seen in the graph as the scattering of points about the line.

EXAMPLE 2. The data show the relationship between the number of students and the number of instructors at a sample of 8 Bay Area community colleges during a recent term.

	X = Number of Students	Y = Number of Faculty
De Anza	26173	846
Foothill	20919	618
West Valley	13800	433
Mission	12814	411
San Jose City	11513	436
Evergreen	10936	330
Gavilan	9092	234
Cabrillo	16369	618

Community College Student Enrollment and Faculty

Find the correlation coefficient _____ = _____ and coefficient of determination _____ = _____

Write the interpretation of the coefficient of determination in the context of the problem.

EXAMPLE 8. Lisa's Lunch Restaurant serves soup and salad. Lisa believes that sales of soup (in dollars of revenue) depends on the temperature. She sells more soup when the weather is cold than when its warmer

Here is the data for a sample of 8 days relating the high temperature for the day with the sales of soup.
X = *High Temperature for the day In degrees F*   Y = *Soup Sales for the day in dollars*

X = temperature	35	49	36	54	43	45	72	65	55	29
Y = soup sales	976	844	820	724	676	880	436	364	472	1060

1. Use STATPLOT on the calculator to create a scatterplot that relates x= temperature to y = sales of soup.

   Does a line look like a reasonable model for this data?

2. Find the coefficient of determination and fill in the sentences for the interpretation:

   ____% of the variation in soup sales is explained by variation in temperature using the regression line.

   ____% of the variation in soup sales can NOT be explained by the variation in temperature using the regression line, but is due to other factors or randomness.

3. Find the correlation coefficient: ____ = _____

   The sign of r is _____ because soup sales _____
   when the temperature increases.

EXAMPLE 9. Suppose that at Ike's Ice Cream Palace, the correlation coefficient for y = revenue in dollars from ice cream sales, and x = high temperature for the day has a correlation coefficient of 0.723.

1. Do sales increase or decrease when the temperature increases? _____How do we know that?

2. Find the coefficient of variation and write its interpretation in the context of the problem

4

Hypothesis Test of the Significance of the Correlation Coefficient

EXAMPLE 9: *OpenStax College, Introductory Statistics http://cnx.org/contents/30189442-6998-4686-ac05-ed152b91b9de@17.44*

A statistics instructor examined the relationship between her students' grades on a midterm exam and their grade on the final exam: x =midterm exam score   y =final exam score

- Data were examined for a small sample of 5 students and then for a larger sample of 11 students. Both samples have almost the same correlation coefficient **r**.
- For the smaller sample, changing any one data value could change the position of the linear regression line a lot. For a larger sample, changing any one data value would not change the the line as much because there are more other points to influence the position of the line.
- As a result of the increased sample size, the data from the larger sample appears more reliable

The reliability of the linear relationship depends on the number of data points, n, and on the value of r

Sample of n = 5 students r = 0.6615						Sample of n = 11 students r = 0.663										
x	65	66	75	70	69	x	65	67	71	71	66	75	67	70	71	69	69
y	175	126	198	163	159	y	175	133	185	163	126	198	153	163	159	151	159

n = 5 students
r = 0.662

n = 11 students
r = 0.663

y = a + bx $\beta \neq 0$ 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 =0.4376 **r = 0.6615**	y = a + bx $\beta \neq 0$ 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 =0.43969 **r = 0.66309**

PVALUE Method To Test Significance Of Correlation Coefficient r:

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

ρ = population correlation coefficient (lower-case Greek letter "rho") ρ is the population parameter. ρ is unknown for the whole population	r = sample correlation coefficient r is the sample statistic. r is the best point estimate of ρ. r is known (calculated from sample data)

- **The hypothesis test lets us make a decision about the value of the population correlation coefficient, ρ, based on the sample data.**
- We will decide if ρ is "significantly different from 0" OR "not significantly different from 0"
- **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*)

- **Two Methods: p-value approach** *(done in class, in textbook and in chapter notes)*
  critical value approach *(in textbook– not done in class, not in chapter notes)*

p-value tells us how likely it is that a given sample correlation coefficient, r, will occur if $\rho = 0$
(if there was not any linear relationship between x and y in the population)

If the p-value $< \alpha$, then the sample correlation coefficient r is "far enough away from 0 to:
Reject Null Hypothesis that Ho: $\rho = 0$. Data show strong enough evidence to conclude Ha: $\rho \neq 0$,
- sample correlation coefficient r is significant (significantly different from 0)
- so we believe that the population correlation coefficient ρ 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. The linear relationship in the sample data is strong and reliable enough to indicate that the linear relationship is likely to be true in the population.

We use the regression line to model the data and predict y values only if the following are satisfied:
(1) if the correlation coefficient is significant
   AND
(2) if you verified by looking at the graph that a line looks to be an appropriate fit for the data
   AND
(3) if the x values you are using as the input for the prediction are between (or equal to) the minimum and maximum x values in the observed data.

If the p-value $> \alpha$, then the sample correlation coefficient r is NOT sufficiently "far from 0":
so we Do Not Reject Null Hypothesis that Ho: $\rho = 0$ The data do not show strong enough evidence to conclude Ha: $\rho \neq 0$,
- sample correlation coefficient r is not significant
- population correlation coefficient $\rho = 0$

➤ We can NOT use the line $\hat{y}$ = a + b*x* to estimate (predict) y based on a given *x* value.
   The linear relationship in the sample data is NOT strong and reliable enough to indicate that the linear relationship exists in the population, so we can only use $\hat{y} = \bar{y}$ to estimate all y values.
   ($\bar{y}$ is the average of all y values.)

What your calculator does for you: Test statistic is $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; degrees of freedom = 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

REGRESSION LINE: $\hat{y} = a + bx$

The process of finding the best fit line $\hat{y} = \mathbf{a} + \mathbf{bx}$ is called **linear regression**.

The **line of best fit** $\hat{y} = \mathbf{a} + \mathbf{bx}$ is also called the **least squares regression line** or just **regression line**

♦ **X is the independent variable**: input variable, horizontal variable, "predictor" variable
♦ **Y is the dependent variable**: output variable, vertical variable, "response" variable
♦ $\hat{y}$ **is the value of y that is estimated by the line** for a corresponding value of x
♦ y is used for observed data; $\hat{y}$ is used for the predicted y values.
♦ **b = slope of the line; it is interpreted as the amount of change in y per unit change in x**
♦ **a = y-intercept**; a is interpreted as the value of y when x = 0 if it makes sense for the problem

$$b = r\frac{s_y}{s_x} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}$$

**We use the regression line to model the data and predict y values only if the following are satisfied:**

(1) if the correlation coefficient is significant
AND
(2) if you verified by looking at the graph that a line looks to be an appropriate fit for the data
AND
(3) if the x values you are using as the input for the prediction are between (or equal to) the minimum and maximum x values in the observed data.

EXAMPLE 8 REVISITED

At Lisa's Lunch Restaurant, Lisa believes that sales of soup (in dollars of revenue) depends on the temperature. She sells more soup when the weather is cold than when its warmer

X = *High Temperature for the day In degrees F*
Y = *Soup Sales for the day in dollars*

The data for a sample of 8 days relating the high temperature for the day with the sales of soup are shown below:

y = soup sales ($)
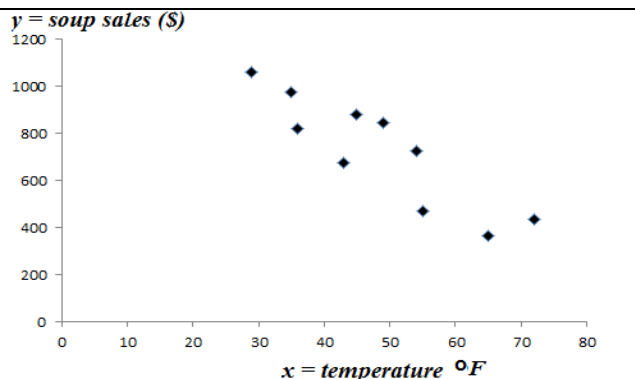
x = temperature °F

X = temperature	35	49	36	54	43	45	72	65	55	29
Y = soup sales	976	844	820	724	676	880	436	364	472	1060

1. Does it appear from the scatterplot that a line is a reasonable model for this data?_____

4. Perform a hypothesis test of the significance of the correlation coefficient

Hypotheses: Ho:_____   Ha:_____

r = _____   pvalue = _____   α = _____
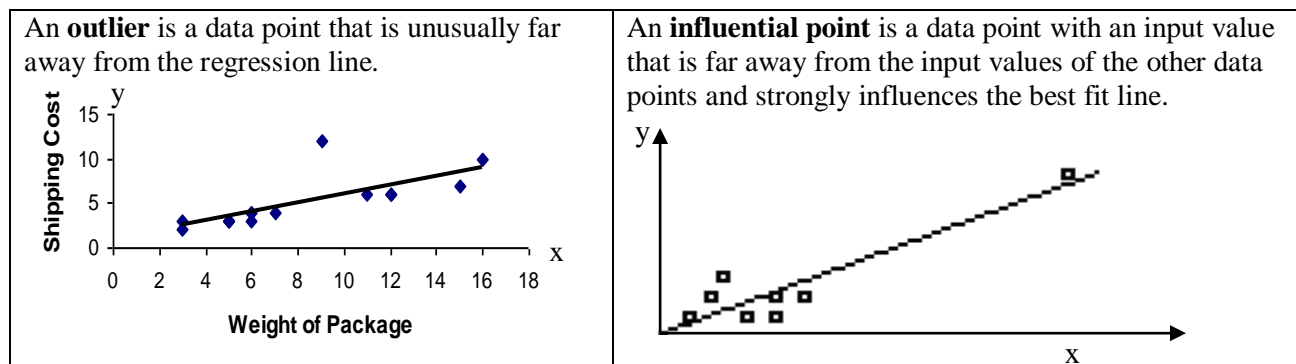
Decision: _____

Conclusion

5. Use the data to find the best-fit line to model how sales of soup relates to and changes with temperature. Use LinRegTTest on your calculator. $\hat{y}$ = _____x + _____

EXAMPLE 8 CONTINUED

6. Graph the best fit line on your scatterplot, using the Y= editor.

7. The best-fit line is also called the regression line. Use the regression line to predict how much soup should Lisa's Lunch Restaurant expects to sell if the temperature is 40 degrees.

8. Suppose that we look at a day when the temperature is 45 degrees.

    a. How much soup does the line predict will be sold? _____

    b. How much soup was actually sold? _____

    c. Is the observed data point for 45 degrees above or below the line? _____

    d. Did the line overestimate or underestimate the amount of soup sold? _____

    e. The residual (or "error") is the difference, between the amount sold and the amount the line predicted.

        Residual $= y - \hat{y} =$ _____

9. Suppose that we look at a day when the temperature is 55 degrees.

    a. How much soup does the line predict will be sold? _____

    b. How much soup was actually sold? _____

    c. Is the observed data point for 45 degrees above or below the line? _____

    d. Did the line overestimate or underestimate the amount of soup sold? _____

    e. The residual (or "error") is the difference, between the amount sold and the amount the line predicted.

        Residual $= y - \hat{y} =$ _____

10. Use the slope to complete the interpretation.
    (In general, you will be required to write the entire interpretation, rather than filling in blanks.)

    Soup sales _____ by _____ for every 1 degree increase in temperature.
            (increase or decrease)   *(value)*

11. Suppose Lisa wants to predict the sales of soup if it were 0 degrees outside. That's very cold.

    a. Should we use the line to predict the sales of soup at a temperature of 0 degrees.

    b. Give some reasons why it might not be appropriate to use the line to predict the sales of soup at a
       temperature of 0 degrees.

12. Suppose Lisa wants to predict the sales of soup if it were 100 degrees outside.

    a. Should we use the line to predict the sales of soup at a temperature of 0 degrees?

    b. Give some reasons why it might not be appropriate to use the line to predict the sales of soup at a
       temperature of 100 degrees.

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 with an input value that is far away from the input values of the other data points and strongly influences the best fit 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 it from the data.

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

TI-83,84:
Use **LinRegTTest** to find a, b, and s to use in the equations below
Press **Y=** key to access the graphing equation editor:
Enter Regression Line: **Y1 = a + bX**
Enter extra lines   **Y2 = a + bX – 2s**
                   Y3 = a + bX + 2s
Make sure your scatterplot is set up and turned on
ZOOM 9 to graph the points and the line
Use **TRACE** to move along the points to determine the (x,y) coordinates of any outliers.
For this class we'll call lines Y2 and Y3 "fence lines" although that is not an official or generally accepted term. Using that terminology will help us understand the concept of outliers in linear regression.

Note: The textbook uses a calculation with 1.9s to determine outliers.
We'll use 2s when doing it graphically.

A data point that is further away from the line of regression than the "fence lines" is an outlier.

EXAMPLE 10:
We are interested in the relationship between the weights of packages and the shipping costs for packages shipped by the Speedy Delivery Co.

x = weight of package (pounds)	5	5	16	9	6	15	7	3	12	6	5	3	12	6	11
y = shipping cost ($)	3	3	10	12	4	7	4	2	6	3	3	3	6	4	6

Identify the (x,y) coordinates of any points in the data that are outliers.

What makes a line be a best fit line? ----Least Squares Criteria for the Best Fit Line

Best fit line $\hat{y} = a + bx$**:** called the **least squares regression line , regression line,** or **line of best fit** .
We use technology to find the values of a (the y intercept) and b (the slope)

The formulas for the best fit line are: $b = r\dfrac{s_y}{s_x} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ and $a = \bar{y} - b\bar{x}$

where S_y is the standard deviation of the y values and S_x is the standard deviation of the x values,
and $\bar{y}$ is the average of the y values and $\bar{x}$ is the average of the x values.

These formulas for the best fit line are developed from optimization techniques in multivariable calculus.
There are some alternative representations of these formulas that look different but are algebraically
equivalent. The calculations can be time consuming and tedious to do by hand.

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 SSE $= \Sigma(y-\hat{y})^2$ **is minimized.**

SSE is the sum of the squares of the residuals, also called <u>**Sum of the Squared Errors.**</u>
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 SSE $= \Sigma(y-\hat{y})^2$ larger than the SSE $= \Sigma(y-\hat{y})^2$ for the best fit line.
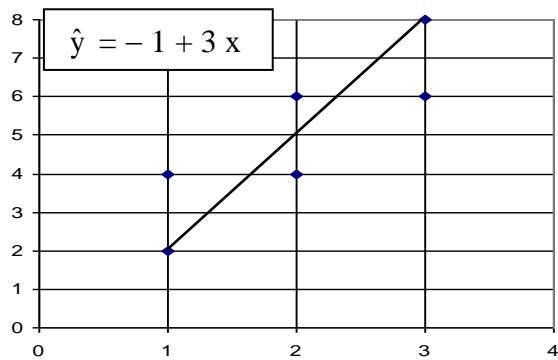
EXAMPLE 12: Both graphs show the same 6 data points but show different lines. One is the best fit
line. Complete the tables to compare the SSE's and choose the least squares regression line.
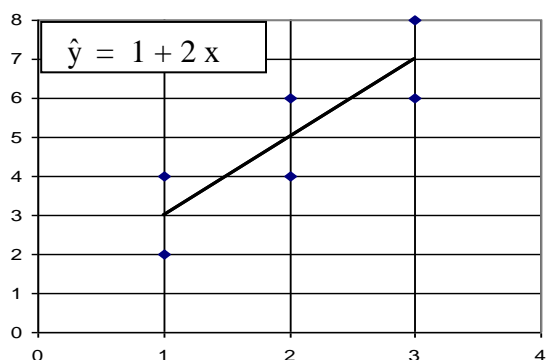
x	y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
1	2	3	-1	1
1	4	3	1	1
2	6	5	1	1
2	4	5	-1	1
3	8	7	1	1
3	6	7	-1	1

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

x	y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
1	2			
1	4			
2	6			
2	4			
3	8			
3	6			

Add (y - $\hat{y}$)² column: SSE $=\Sigma(y-\hat{y})^2=$ _____

Line _____ is the best fit according to the _____ criteria

because _____ .

10

CALCULATOR INSTRUCTIONS: TI-83, 83+, 84+:

DRAWING A SCATTERPLOT **TI-83, 83+, 84:** **2nd STATPLOT 1**	On Off Type *Highlight the scatterplot icon* $\vdots$ *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 **:** $\neq 0$ **<0**   **>0** *Highlight* $\neq 0$ *ENTER*
RegEQ: *Leave RegEq blank*
Calculate *Highlight Calculate ; then press ENTER*

LinRegTTest OUTPUT SCREEN

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

GRAPH THE BEST FIT LINE ON SCATTER PLOT using equation found with the LinRegTTest:

Find equation of line $\hat{\mathbf{y}} = \mathbf{a} + \mathbf{b}x$ 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).

Press ZOOM → 9:ZoomStat.

IDENTIFY OUTLIERS

(Note: your text book uses the term "potential outliers".)

Graph 3 lines on the same screen as the scatterplot.
Y1 = a+ bx
Y2 = a+bx–2s
Y3 = a+bx+2s

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, you can zoom in to see it better or you can do the calculation numerically to determine if it is outside or inside the lines.

CHECKLIST: 10 SKILLS AND CONCEPTS YOU NEED TO LEARN IN CHAPTER 12

1. Identify which variable is independent and which variable is dependent, from the context (words) of the problem.

2. Know calculator skills for items 3, 4, 5, 6, 9 below.
    Complete calculator instructions are near the end of these notes and will be demonstrated in class.

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 <u>correlation coefficient</u> to determine if a <u>significant linear relationship</u> exists and to assess the strength of the linear relationship (hypothesis test of significance of r using the p-value approach).

5. Find and interpret the <u>coefficient of determination</u> to determine
    a) what percent of the variation in the dependent variable is explained by the variation in the independent variable using the best fit line,
    b) what percent of the variation in the dependent variable is not explained by the line
    What does the scattering of the points about the line represent?

6. Find and use the <u>least squares regression line</u> to model and explore the relationship between the variables, finding predicted values within the domain of the original data, 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.
    In order to use the line to predict, ALL of the following conditions must be satisfied:
    a) scatterplot of data must be well modeled with a line – visually check the graph to observe if a line is a reasonable fit to the data
    b) p-value $< \alpha$
    c) 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 a <u>verbal interpretation of the slope</u> as marginal change in context of the problem.
    (Marginal change is change in x per unit of y, stated in the words of and using the numbers and units of the particular problem.)
9. Understand the importance of outliers and influential points

10. Understand the concept of the <u>least squares criteria</u> for determining the best fit line.

SKILLS PRACTICE 1 :

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.

	X = Number of Students	Y = Number of Faculty
De Anza	26173	846
Foothill	20919	618
West Valley	13800	433
Mission	12814	411
San Jose City	11513	436
Evergreen	10936	330
Gavilan	9092	234
Cabrillo	16369	618

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. a) How many faculty are predicted for a college with 11,513 students?
   b) What is the residual ($y - \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 11,513?
   c) Did value predicted by the line overestimate or underestimate the observed value?

9. a) How many faculty are predicted for a college with 20,919 students?
   b) What is the residual ($y - \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 20,919?
   c) 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:

    a) 4000 students? _____   b) 14000 students? _____   c) 40000 students? _____
    Explain why or why not.

13

SKILLS PRACTICE 2

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

GreatBuy Electronics Store is 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 = price of this DVD during a one week period
y = number of this DVD sold during a one week period

x ($)	y (DVDs)
13	370
15	400
15	330
16	380
18	250
13	340
16	350
15	310
16	360
18	260
13	380
18	290

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. a) Predict the sales when the price is $15.
   b) What is the residual (y $- \hat{y}$: difference between the observed y and predicted $\hat{y}$) when x = 15?
   c) Did value predicted by the line overestimate or underestimate the observed value?

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?

SKILLS PRACTICE 3

Is there a relationship between the number of absences a student has during the quarter (out of 54 class sessions) and the grade the student earns for the course?

Days Absent	0	0	1	1	2	2	3	4	4	5	8	9	9	10	12	12	15	16	20	22
Course Grade	3	4	2	1	4	2	2	2	4	3	3	2	1	3	0	2	1	2	1	0

1. Identify the independent variable: ___ = _____

2. Identify the dependent variable: ____ = _____

3. Find the best fit line and write the equation of the line. _____

4. Graph a scatterplot of the data, showing the best fit line

5. Find the correlation coefficient and the coefficient of determination:

6. Test the significance of the correlation coefficient. Show your work and clearly state your conclusion

7. Write the interpretation of the coefficient of determination in the context of the data.

8. Write the interpretation of the slope of the regression line, in the context of the data.

9. a) Predict the grade for a student with 5 absences. _____

    b) Find the residual y $- \hat{y}$ (difference between observed y and predicted $\hat{y}$) when x = 5: _____

    c) Did value predicted by the line overestimate or underestimate the observed value?

10. a) Predict the grade for a student with 15 absences. _____

    b) Find the residual y $- \hat{y}$ (difference between observed y and predicted $\hat{y}$) when x = 15: _____

    c) Did value predicted by the line overestimate or underestimate the observed value?

11. a) Predict the grade for a student with 9 absences. _____

    b) How do the data compare to the predicted value?

12. Predict the grade for a student with 40 absences. Explain why the best fit line predicts a grade that does not make any sense in this problem.

SKILLS PRACTICE 4

The population of River City is recorded by the U.S. Census every 10 years.
Between the census in 1950 and 2010, the population has more than doubled.
The population data for the past 7 censuses are:

Year	1950	1960	1970	1980	1990	2000	2010
Population	22, 250	23,100	26,250	30, 200	35,250	41,300	52,100

1. Find the correlation coefficient and conduct a hypothesis test for significance.

2. Graph a scatterplot of the data and graph the best fit line to see how the data fit the line.

3. Does the best fit line appear to be good model for this data? Explain.

SKILLS PRACTICE 5 :

We are interested in the relationship between the weights of packages and the shipping costs for packages shipped by the Speedy Delivery Co.

x = weight of package (pounds)	5	5	16	9	6	15	7	3	12	6	5	3	12	6	11
y = shipping cost ($)	3	3	10	12	4	7	4	2	6	3	3	3	6	4	6

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. Find the correlation coefficient and conduct a hypothesis test for significance.

5. Find the coefficient of determination and write its interpretation 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 shipping cost for a package that weighs 10 pounds:

SKILLS PRACTICE 6 : EXERCISE 12.87   Name _____

This has been modified from the above content in the textbook.

The table shows average heights for American boys.
(Source:Physician's Handbook, 1990)

X = Age (years)	Y = Height (cm)
0	30.8
2	83.8
3	91.4
5	106.6
7	119.3
10	137.1
14	157.5

a. Calculate the least squares line: _____

b. Find the correlation coefficient and write the hypothesis test to determine if the correlation coefficient is significant.

c. Find the estimated average height for a 12 year–old. _____

d. Find the estimated average height for 5 year old. _____

Is the observed data point above or below the line? _____

Does the line overestimate or underestimate the actual average height for a 5 year old? _____

Calculate the residual (observed– predicted) : $y - \hat{y}$ = _____

e. Use the least squares line to estimate the height for a 50 year–old man. Is this answer is reasonable? Why or why not? *(You may want to change units to feet and inches to understand this number better. Height is given in cm ; to convert to inches, divide by 2.54; then to convert to feet, divide by 12)*

f. What is the slope of the least squares (best-fit) line? Write a sentence that interprets both the sign and the numerical value of the slope as a rate of change in the context of this problem.

g Make a scatter plot of the data and graph the best fit line. Examine each point to see how close or far away it is from the line. Does it appear that a line is the best way to fit the data? Why or why not?

h. Redo the graph, line, and correlation coefficient using ages 2 through 14 but not the height at birth age 0. Does the new line fit the data points excluding age 0 better than the old line fit with the age 0 data included?

i. Compare the new correlation coefficient to the correlation coefficient in part (e).

Data with ages 0 through 14: r =   _____   Data with ages 2 through 14: r = _____

Which indicates stronger linear correlation? _____

j. Write the interpretation of the coefficient of determination for the data using ages 2 through 14 only.