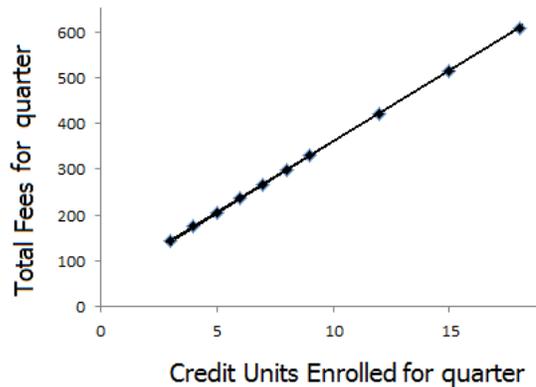


## Chapter 12 : Linear Correlation and Linear Regression

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.

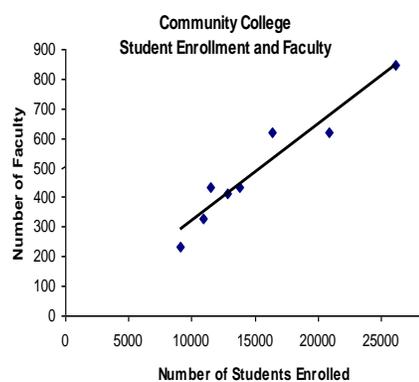
At a community college students pay a basic fee of \$50 per quarter, plus a fee of \$31 per credit unit:



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.

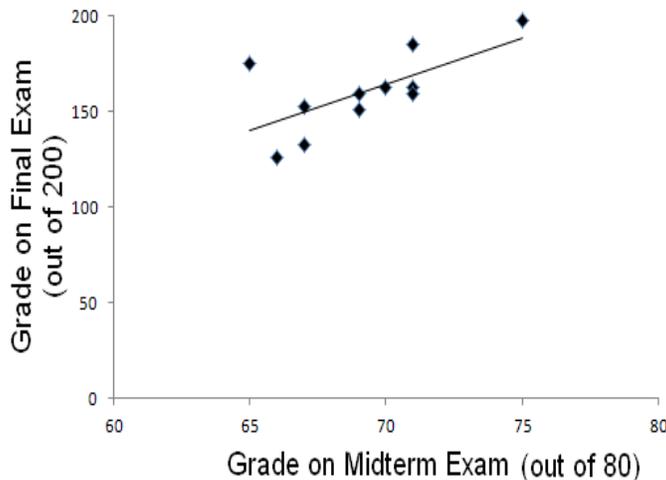
Relationship between number of students and number of instructors at a sample of 8 Bay Area community colleges during a recent term.



	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

Linear Regression and Correlation Notes, by Roberta Bloom, De Anza College

This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).



- Some material, including Example 3, is derived and remixed from **Introductory Statistics** from Open Stax (Illowsky/Dean) available for download free at <http://cnx.org/content/11562/latest/> or <https://openstax.org/details/introductory-statistics>
- Some material, including Example 9, is derived and remixed from **Inferential Statistics and Probability: A Holistic Approach**, by Maurice Geraghty, De Anza College, 1/1/2018, <http://professormo.com/holistic/HolisticStatisticsRev190204.pdf>

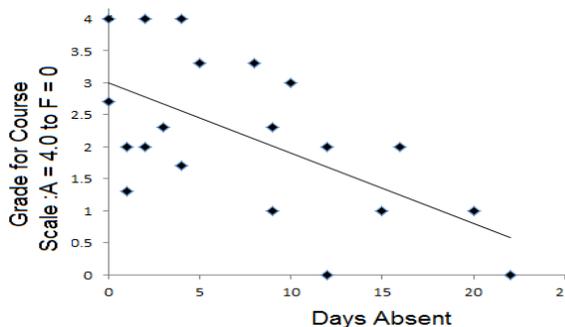
**EXAMPLE 4.**

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)

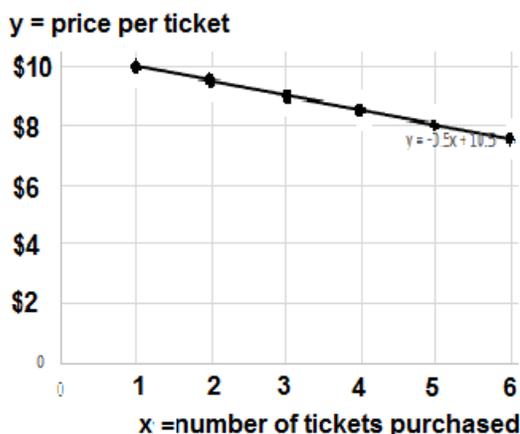


Days Absent	0	0	1	1	2	2	3	4	4	5	8	9	9	10	12	12	15	16	20	22
Course Grade	3	4	2	1	4	2	2	2	4	3	3	2	1	3	0	2	1	2	1	0

**EXAMPLE 5.**

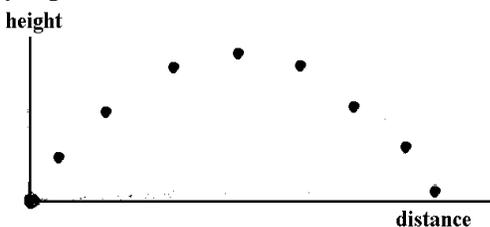
A sightseeing tour bus charges \$10 per ticket.

Their "Family Tour" plan offers discounts per ticket that depend on the total number of tickets purchased, up to 6 tickets.



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

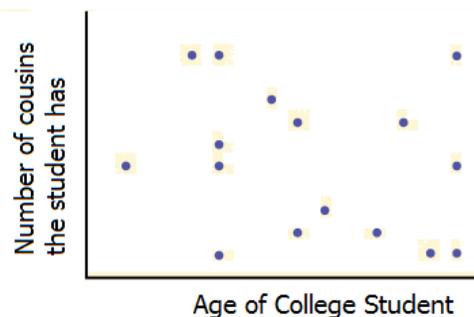
**EXAMPLE 6.** A golf ball is hit into the air from the ground. Its height above ground (y) and the horizontal distance (x) it has traveled are related by a parabolic curve.



**EXAMPLE 7.**

X = the age of a college student

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 ?

## MEASURING HOW WELL A LINEAR MODEL FITS THE DATA

### CORRELATION COEFFICIENT $r$ :

A number that measures the strength of the linear relationship between two quantitative variables.

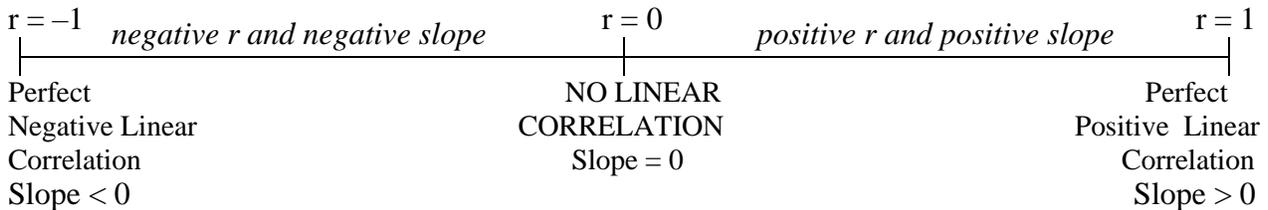
Symbol:  $r$

Values:  $-1 \leq r \leq 1$

- If all points lie exactly on the line, the correlation coefficient is  $r = +1$  or  $r = -1$ .
- 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  $0$ .
- If there is no linear relationship between the variables, the correlation coefficient is  $r = 0$

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  $r < 0$



Formulas for Correlation Coefficient used by your calculator

Conceptually  $r$  examines the variation in  $x$  and  $y$  jointly (numerator) compared to the variation in each variable separately (denominator)

Theoretical formula defining  $r$

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

“Easier” formula for doing calculations

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

**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) :

$r = -1.0$ ,  $r = 0.66$ ,  $r = -0.61$ ,  $r = 0.96$ ,  $r = 1.0$ ,  $r = 0$

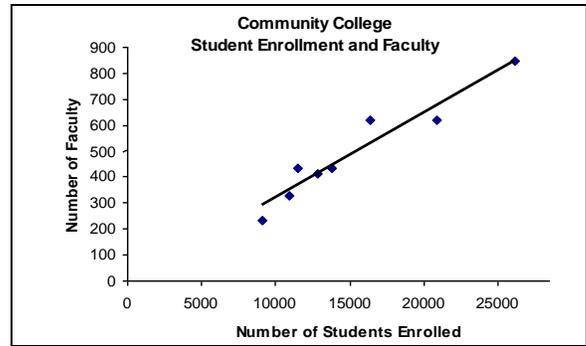
**For each graph determine which of value of  $r$  above best corresponds to the graph and write the value of  $r$  on the graph in the form “ $r = \underline{\quad}$ ”.**

### 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



Find the correlation coefficient \_\_\_\_\_ = \_\_\_\_\_ and coefficient of determination \_\_\_\_\_ = \_\_\_\_\_

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

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

Here is data for a sample of 10 days relating high temperature for the day with sales revenue of soup.  
 X = High Temperature for the day in degrees F    Y = Soup Sales Revenue for the day in dollars

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

1. Use STATPLOT on the calculator to create a scatterplot for x= temperature to y = soup sales revenue.

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 revenue from soup sales is explained by variation in temperature using the regression line.

\_\_\_\_\_ % of the variation in revenue from 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 revenue from soup sales

\_\_\_\_\_ when the temperature increases.

## Hypothesis Test of the Significance of the Correlation Coefficient

What is a “good” value for the correlation coefficient?

**EXAMPLE 9:** Source: Inferential Statistics and Probability: A Holistic Approach, Maurice Geraghty De Anza College; 1/1/2018 <http://professormo.com/holistic/HolisticStatisticsRev190204.pdf> This example is used and adapted under Creative Commons Attribution-ShareAlike 4.0 License.

**16 student volunteers drank a randomly assigned number of cans of beer. Thirty minutes later a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.**

**x = number of beers                      y = blood alcohol content (BAC)**

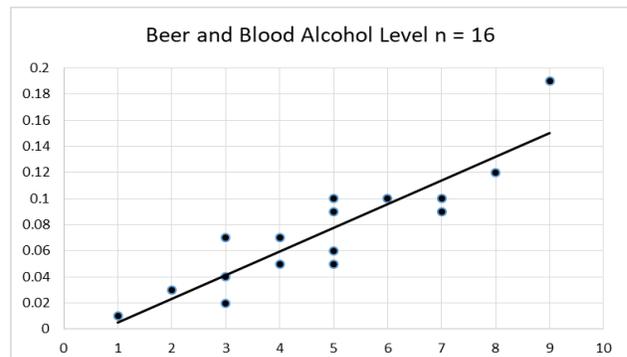
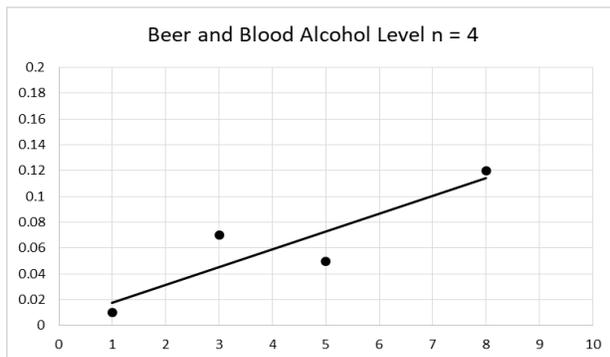
- Data were also examined for a smaller sample of 4 students (subsample taken from the 16 students).
- Data from the larger sample appears more reliable
  - Best fit line for smaller sample is very sensitive to change in one data value due to small sample size.
  - Best fit line for the larger sample is less sensitive to change in any one data value because there are more other points to influence the position of the line.

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

<b>Sample of n = 4 students</b>	x	1	5	3	8
	y	.01	.05	.07	.12

<b>r = 0.897 for both samples of data</b>
---

<b>Sample of n = 16 students</b>	x	5	2	9	8	3	7	3	5	3	5	4	6	5	7	1	4
	y	.1	.03	.19	.12	.04	.1	.07	.06	.02	.05	.07	.1	.09	.09	.01	.05



$y = a + bx$ $\beta \neq 0$ and $\rho \neq 0$ <b>t=2.87</b> <b>p = 0.103</b> df = 2 (df = n-2 = 4-2) a = .004 b = .014 s = 0.025 $r^2 = 0.805$ <b>r = 0.897</b>
--

$y = a + bx$ $\beta \neq 0$ and $\rho \neq 0$ <b>t=7.59</b> <b>p = 0.0000025</b> df = 14 (df = n-2 = 16-2) a = -.01 b = .018 s = 0.02 $r^2 = 0.804$ <b>r = 0.897</b>
---

## Reference Notes for PVALUE Method To Test Significance Of Correlation Coefficient r

Objective: Determine if the linear relationship in the sample data is strong and reliable enough to use it as an estimate of the model for a linear relationship for the whole population.

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

- The hypothesis test lets us make a decision about the value of the population correlation coefficient,  $\rho$ , based on the sample data.
- Decide if  $\rho$  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 p-value  $> \alpha$ , then the sample correlation coefficient  $r$  is NOT sufficiently "far from 0":**

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
- we assume that the population correlation coefficient  $\rho = 0$

➤ We can NOT use the line  $\hat{y} = a + bx$  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.)

**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  $\rho$  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.

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.**

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

**LINEAR REGRESSION: FINDING THE BEST FIT LINE:  $\hat{y} = a + bx$**

Line of best fit  $\hat{y} = a + bx$  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

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

We use the regression line to model the relationship between the variables and to predict y values only if all the following are satisfied:

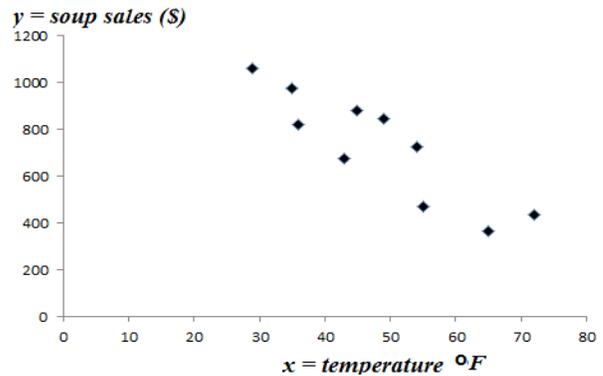
- (1) if the correlation coefficient is significant
- AND**
- (2) if you verified by looking at the graph that a line looks to be an appropriate fit for the data
- AND**
- (3) if the x values you are using as the input for the prediction are between (or equal to) the minimum and maximum x values in the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

**EXAMPLE 8 REVISITED**

At Lisa’s Lunch Restaurant, Lisa believes that revenue (in dollars) from sales of soup 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 Revenue for the day in dollars

The data for a sample of 10 days relating the high temperature for the day with the revenue from sales of soup are shown below:



X = temperature	35	49	36	54	43	45	72	65	55	29
Y = revenue from 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 = \_\_\_\_\_  $\alpha$  = \_\_\_\_\_

Decision: \_\_\_\_\_

Conclusion

5. Use the data to find the best-fit line to model how revenue from soup sales changes with temperature. Use LinRegTTest on your calculator. Graph the best fit line on your scatterplot, using the Y= editor.

$$\hat{y} = \text{_____} + \text{_____} x$$

**EXAMPLE 8 CONTINUED**

6. The best-fit line is also called the regression line. Use the regression line to predict how much revenue (income) Lisa’s Lunch Restaurant expects from soup sales if the temperature is 40 degrees.
7. Suppose that we look at a day when the temperature is 45 degrees.
- How much revenue from soup sales is predicted by the line? \_\_\_\_\_
  - What was the actual revenue from soup sales? \_\_\_\_\_
  - Is the observed data point for 45 degrees above or below the line? \_\_\_\_\_
  - Did the line overestimate or underestimate the revenue from soup sales? \_\_\_\_\_
  - The residual (or error) = “ actual observed revenue” MINUS “predicted revenue” from soup sales.  
$$\text{Residual} = y - \hat{y} = \underline{\hspace{2cm}}$$

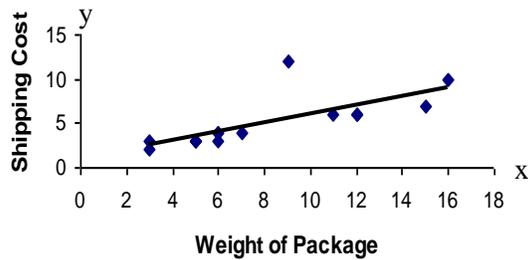
(The residual  $y - \hat{y}$  is the vertical distance that the point is above or below the line in the graph.)
8. Suppose that we look at a day when the temperature is 55 degrees.
- How much revenue from soup sales is predicted by the line? \_\_\_\_\_
  - What was the actual revenue from soup sales? \_\_\_\_\_
  - Is the observed data point for 55 degrees above or below the line? \_\_\_\_\_
  - Did the line overestimate or underestimate the revenue from soup sales? \_\_\_\_\_
  - The residual (or error) = “ actual observed revenue” MINUS “predicted revenue” from soup sales.  
$$\text{Residual} = y - \hat{y} = \underline{\hspace{2cm}}$$

(The residual  $y - \hat{y}$  is the vertical distance that the point is above or below the line in the graph.)
9. 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 revenue \_\_\_\_\_ by \_\_\_\_\_ for every 1 degree increase in temperature.  
*(increase or decrease) (value)*
10. Suppose Lisa wants to predict the sales of soup if it were 0 degrees outside. That’s very cold.
- Should we use the line to predict the sales of soup at a temperature of 0 degrees?
  - 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.
11. Suppose Lisa wants to predict the sales of soup if it were 100 degrees outside.
- Should we use the line to predict the sales of soup at a temperature of 100 degrees?
  - 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.

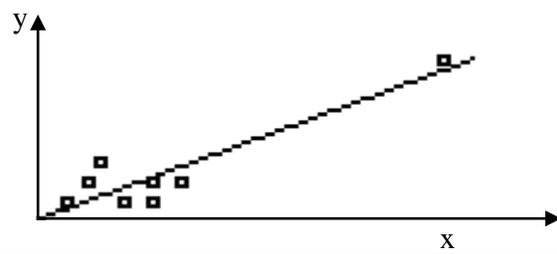


## OUTLIERS IN LINEAR REGRESSION

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:** A graphical way to identify outliers in a scatterplot  
 Use **LinRegTTest** to find a, b, and s  
 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: STAT** to graph the points and the line  
 Use **TRACE** to move among the points to find (x,y) coordinates of outliers.

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

**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

$y = a + bx$ $\beta \neq 0$ and $\rho \neq 0$ $t = 4.0288$ $p = 0.00143$ $df = 13$ $a = 0.99$ $b = 0.51$ $s = 1.97$ $r^2 = 0.555$ $r = 0.745$	$Y1 = .99 + .51 * X$ (this is best fit line that is the center line of the three parallel lines in the graph)  $Y2 = .99 + .51 * X - 2 * 1.97$  $Y3 = .99 + .51 * X + 2 * 1.97$  Set up and turn on scatterplot Then ZOOM 9:Stat	<p>A scatterplot with 'Weight of Package' on the x-axis (0 to 16) and 'Shipping Cost' on the y-axis (0 to 15). A regression line is shown, along with two parallel lines above and below it representing a confidence interval. The outlier at (9, 12) is clearly above the upper confidence interval line.</p>
--	---	---

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:  $b = r \frac{s_y}{s_x} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$  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, and can also be derived using linear algebra.

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.

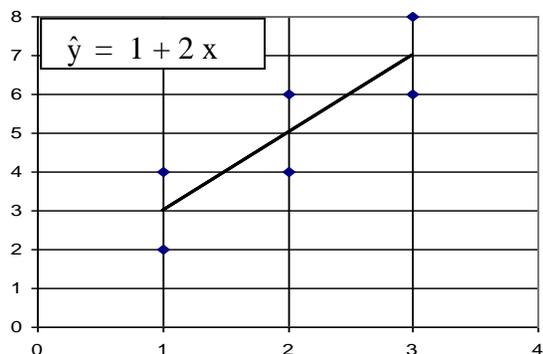
**Definition of Best Fit Line:** The best fit line is the line for which  $SSE = \sum (y - \hat{y})^2$  is minimized.

**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

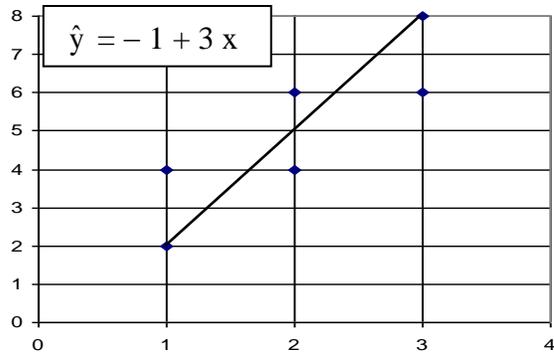
Any other line that you might try to fit through these points will have the sum of the squared residuals  $SSE = \sum (y - \hat{y})^2$  larger than the  $SSE = \sum (y - \hat{y})^2$  for the best fit line.

**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  
 $SSE = \sum (y - \hat{y})^2 = 6$



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:  $SSE = \sum (y - \hat{y})^2 =$  \_\_\_\_\_

Line \_\_\_\_\_ is the best fit according to the \_\_\_\_\_ criteria

because \_\_\_\_\_.

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

### **DRAWING A SCATTERPLOT**

**TI-83, 83+, 84+:**  
**2<sup>nd</sup> 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*

### **LinRegTTest OUTPUT SCREEN**

**LinRegTTest**

**y = a + bx**

**β ≠ 0 and ρ ≠ 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<sup>2</sup> = coefficient of determination**

**r = correlation coefficient**

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

Find equation of line  $\hat{y} = a + bx$   
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).

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 correlation coefficient to determine if a significant linear relationship 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 coefficient of determination 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 least squares regression line 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)  $p\text{-value} < \alpha$
  - 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 verbal interpretation of the slope as marginal change in context of the problem. (Marginal change is change in  $y$  per unit of change in  $x$ , stated in the words of and using the numbers and units of the particular problem. See examples done in class and see textbook for how to write this interpretation.)
9. Understand the importance of outliers and influential points
10. Understand the concept of the least squares criteria for determining the best fit line.