CHAPTER 12 LINEAR REGRESSION AND CORRELATION

INTRODUCTION

Professionals often want to know how two or more numeric variables are related.

For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?

In another example, your income maybe determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.

The type of data described in the examples is bivariate data — "bi" for two variables. In reality, statisticians use multivariate data, meaning many variables. In this chapter, you will be studying the simplest form of regression, "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions. You will also study correlation which measures how strong the relationship is.

12.1 | LINEAR EQUATIONS

LINEAR EQUATIONS

Linear regression for two variables is based on a linear equation with one independent variable.

The equation has the form: $y = a+bx$

a and b are constant numbers.

x is the independent variable

y is the dependent variable.

Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable

LINEAR EQUATIONS

The graph of a linear equation of the form $y=a+bx$ is a straight line.

Any line that is not vertical can be described by this equation.

EXAMPLE OF LINEAR EQUATIONS

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is $32 per hour plus a $31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job. Find the equation that expresses the total cost in terms of the number of hours required to complete the job.

Let x= the number of hours it takes to get the job done.

Let y= the total cost to the customer.

The $31.50 is a fixed cost.

If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only.

The total cost is: $y= 31.50 + 32x$

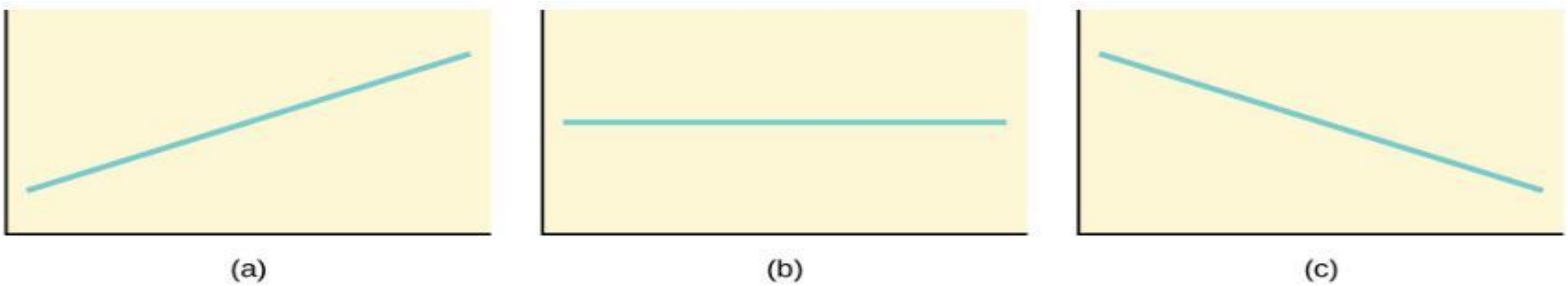
SLOPE AND Y-INTERCEPT OF A LINEAR EQUATION

For the linear equation **y=a+bx**,

b=slope and a=y-intercept.

From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the ycoordinate of the point (0, a) where the line crosses the y-axis.

THREE GRAPHS FOR LINEAR EQUATIONS



Three possible graphs of $y = a + bx$.

(a)If $b > 0$, the line slopes upward to the right.

(b)If $b = 0$, the line is horizontal.

(c) If $b < 0$, the line slopes downward to the right.

EXAMPLE OF LINEAR EQUATIONS

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y= 25 + 15x$.

What are the independent and dependent variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

SOLUTION TO EXAMPLE

The independent variable (x) is the number of hours Svetlana tutors each session.

The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 (a = 25).

At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when x= 0).

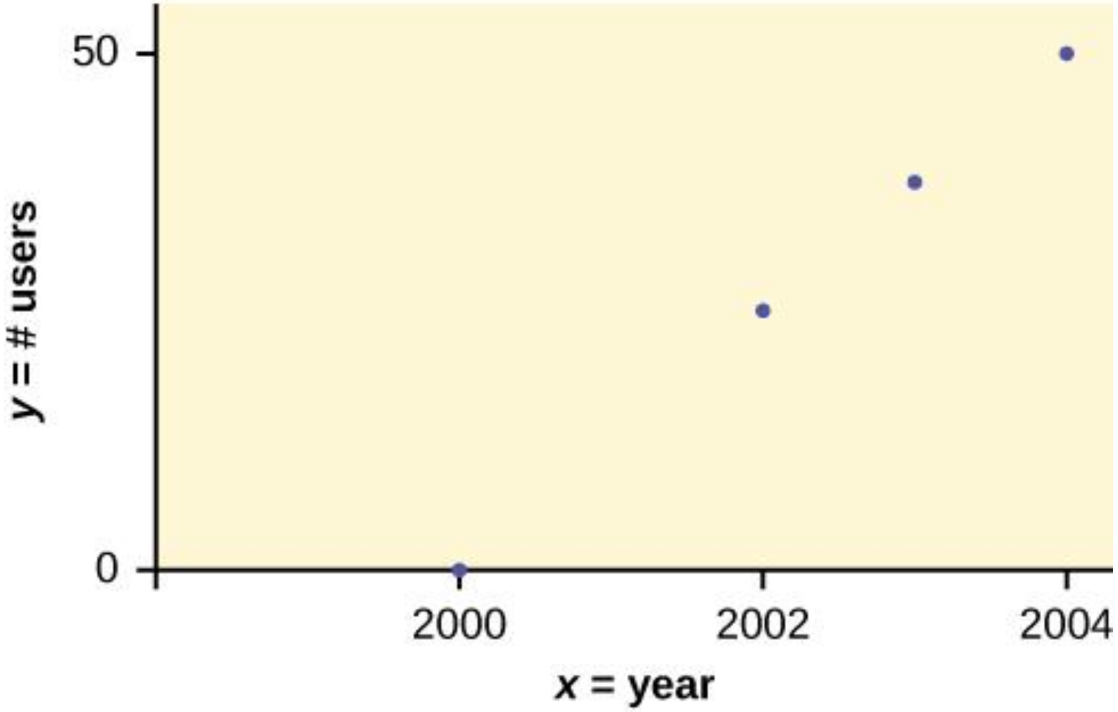
The slope is 15 (b= 15).

For each session, Svetlana earns $15 for each hour she tutors.

12.2 | SCATTER PLOTS

SCATTERPLOT

Before we take up the discussion of linear regression and correlation, we need to examine a way to display the relation between two variables x and y. The most common and easiest way is a scatter plot. The following example illustrates a scatter plot.



SCATTERPLOT

A scatter plot shows the direction of a relationship between the variables.

A clear direction happens when there is either:

- High values of one variable occurring with high values of the other variable or low values of one variable occurring with low values of the other variable.

- High values of one variable occurring with low values of the other variable.

You can determine the strength of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

For a linear relationship there is an exception. Consider a scatter plot where all the points fall on a horizontal line providing a "perfect fit."

The horizontal line would in fact show no relationship.

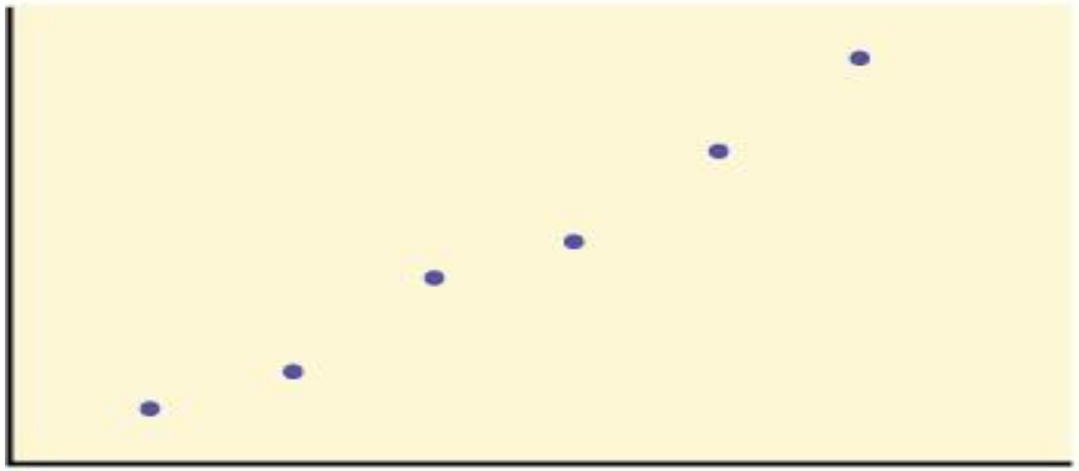
TRENDS IN SCATTERPLOTS

When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.

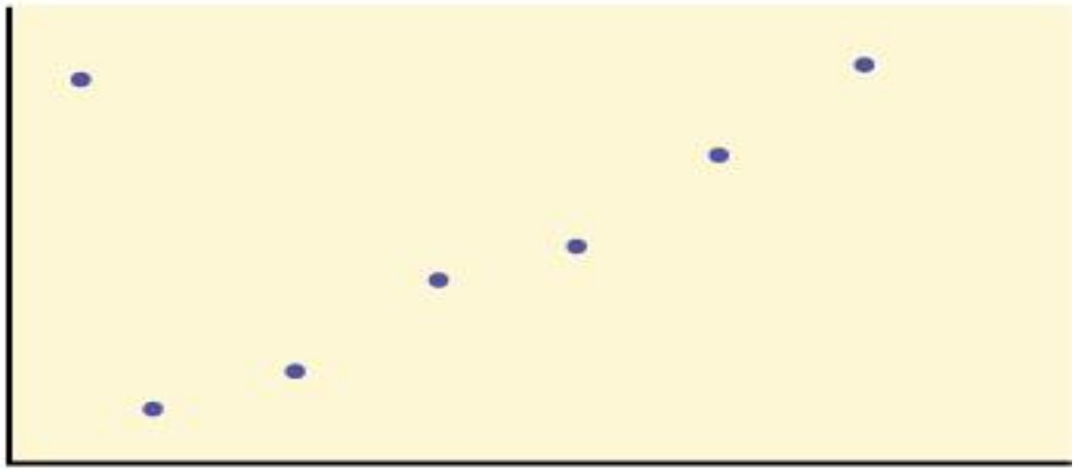
TRENDS IN SCATTERPLOTS

When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.

PATTERNS IN SCATTERPLOTS

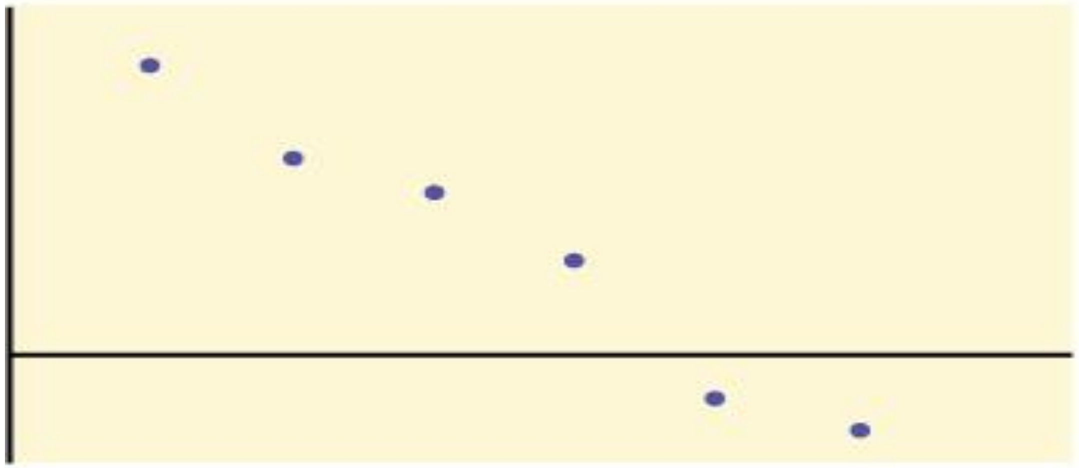


(a) Positive linear pattern (strong)

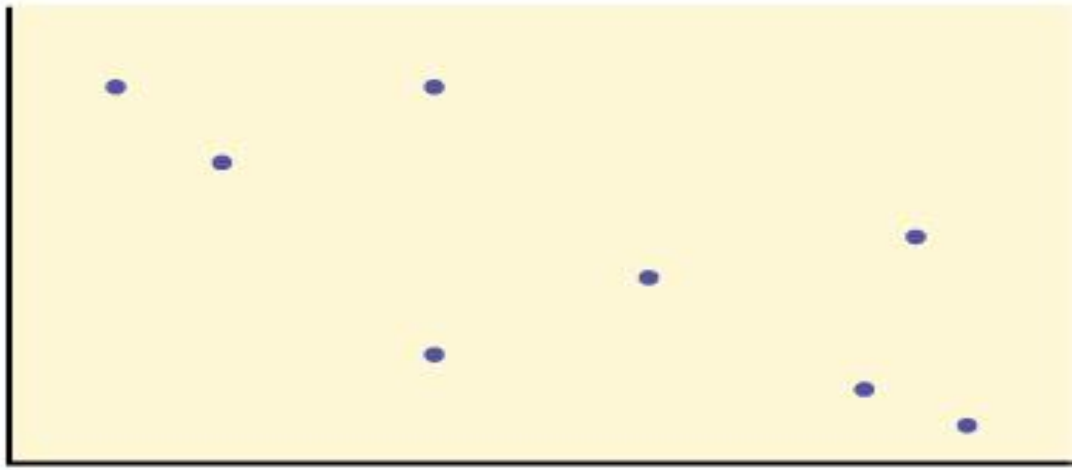


(b) Linear pattern w/ one deviation

PATTERNS IN SCATTERPLOTS

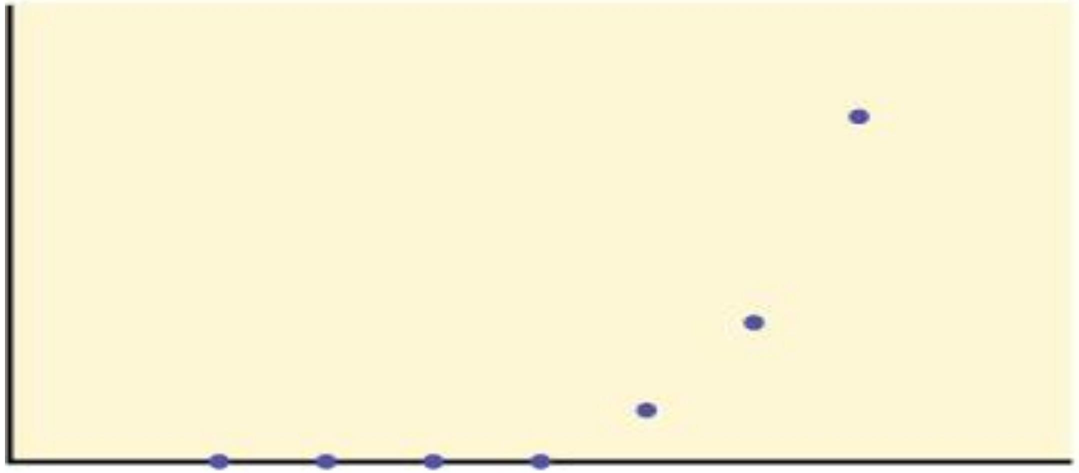


(a) Negative linear pattern (strong)

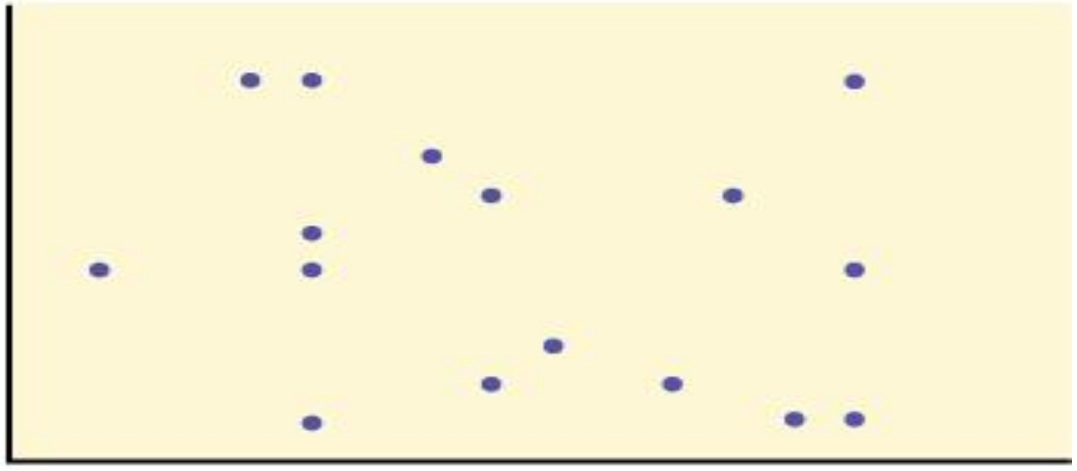


(b) Negative linear pattern (weak)

PATTERNS IN SCATTERPLOTS



(a) Exponential growth pattern



(b) No pattern

SCATTERPLOT TO LINEAR REGRESSION

In this chapter, we are interested in scatter plots that show a linear pattern. Linear patterns are quite common.

The linear relationship is strong if the points are close to a straight line, except in the case of a horizontal line where there is no relationship.

If we think that the points show a linear relationship, we would like to draw a line on the scatter plot. This line can be calculated through a process called **linear regression**.

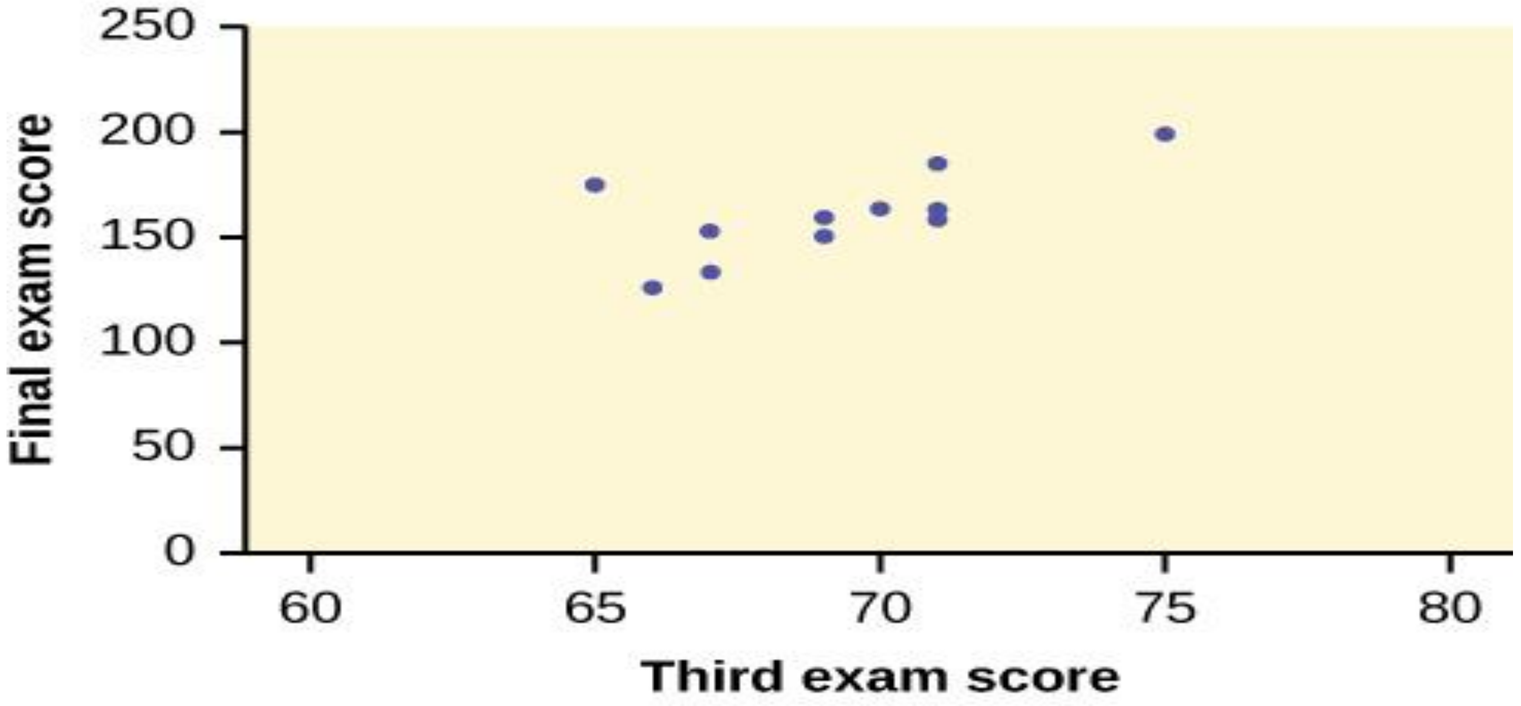
However, we only calculate a regression line if one of the variables helps to explain or predict the other variable. If x is the independent variable and y the dependent variable, then we can use a regression line to predict y for a given value of x.

12.3 | THE REGRESSION EQUATION

LINE OF BEST FIT

Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to "fit“ a straight line. This is called a **Line of Best Fit** or **Least-Squares Line**.

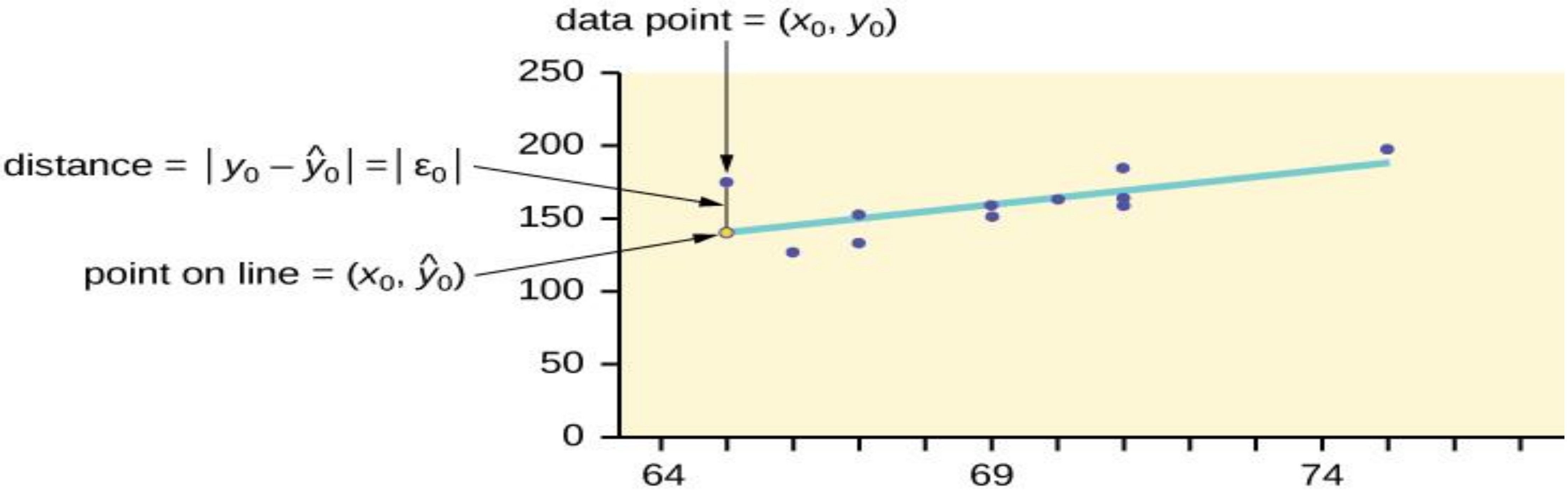
LEAST-SQUARES REGRESSION LINE



The third exam score, x, is the **independent** variable and the final exam score, y, is the **dependent** variable. We will plot a regression line that best "fits" the data.

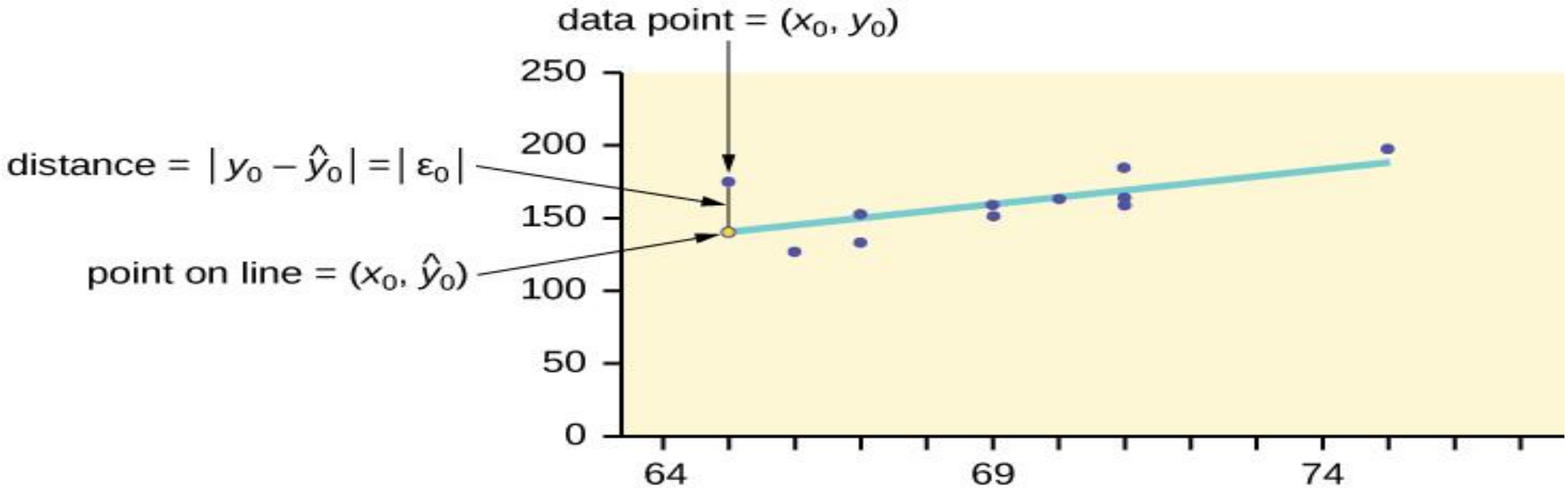
If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a **least-squares regression line** to obtain the best fit line.

Y-HAT



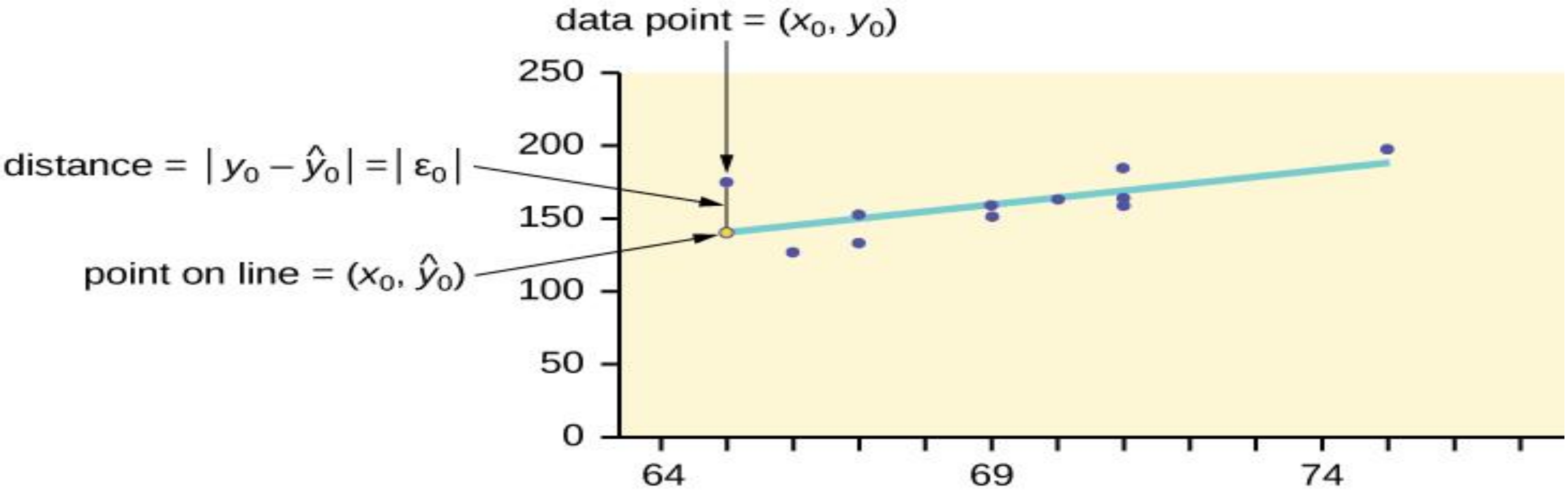
The ŷ is read "yhat“ and is the estimated value of y. It is the value of y obtained using the regression line. It is not generally equal to y from data.

RESIDUAL OR ERROR



The term y0 – ŷ0 = ε0 is called the **"error"** or **residual**. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y. In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

RESIDUAL OR ERROR



If the observed data point lies above the line,the residual is positive, and the line underestimates the actual data value for y. If the observed data point lies below the line, the residual is negative,and the line overestimates that actual data value for y. In the diagram, y0 – ŷ0 = ε0 is the residual for the point shown. Here the point lies above the line and the residual is positive.

SUM OF SQUARED ERRORS (SSE)

ε= the Greek letter epsilon

For each data point, you can calculate the residuals or errors,

yi -ŷi =εi for i= 1, 2, 3, ..., 11.

Each |**ε**| is a vertical distance.

If you square each **ε** and add them together, you get the **Sum of Squared Errors (SSE).**

Using calculus, you can determine the values of a and b that make the SSE a minimum. When you make the SSE a minimum, you have determined the points that are on the line of best fit.

It turns out that the line of best fit has the equation: **y-hat= a + bx**

LINEAR REGRESSION

The process of fitting the best-fit line is called **linear regression**.

The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line.

The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible.

Any other line you might choose would have a higher SSE than the best fit line.

This **best fit line** is called the **least-squares regression line**.

TO CREATE A SCATTERPLOT AND THE LINE OF BEST FIT IN THE CALCULATOR

DATA ENTRY

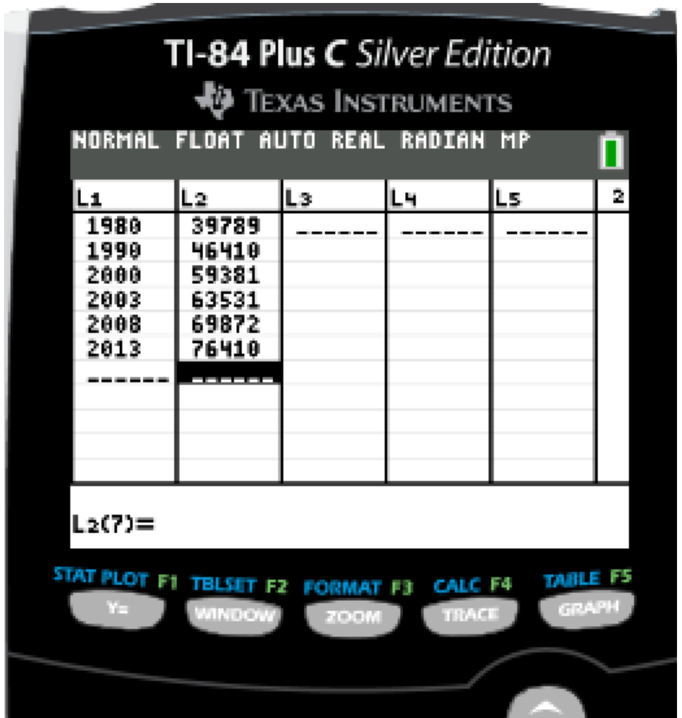
Enter Variable 1 (x—independent--predictor) into list (L1)

Enter Variable 2 (y – dependent—response) into list (L2)

- Be sure to keep the data paired as presented

EXAMPLE: Does a relationship exist between the year and population within 20 minutes of Wal-Mart in Gordon County, GA?

Year	Population within 20 minutes of Wal-Mart
1980	39789
1990	46410
2000	59381
2003	63531
2008	69872
2013	76410



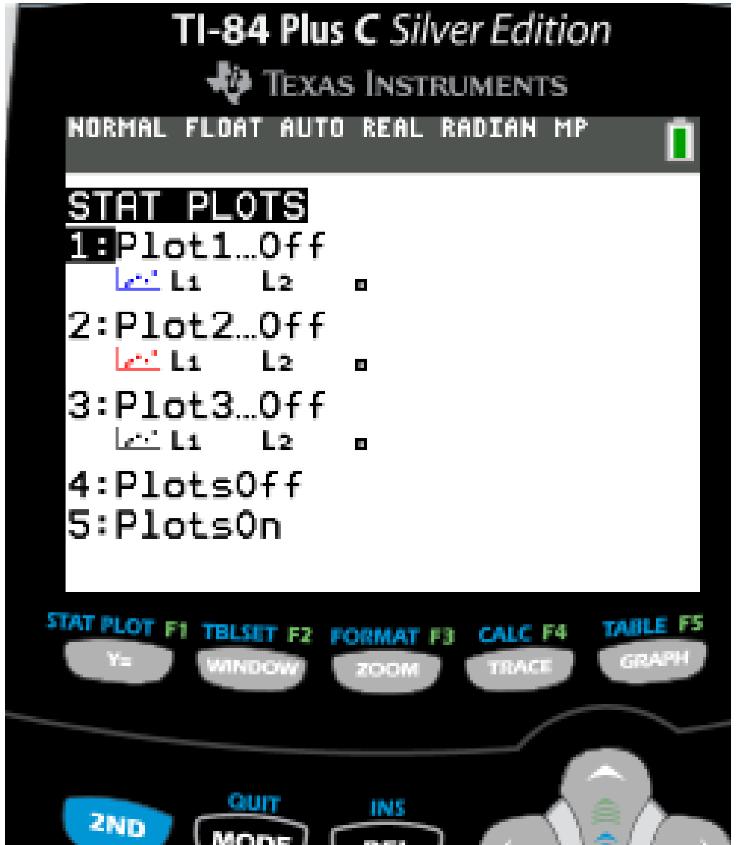
```
NORMAL FLOAT AUTO REAL RADIAN MP
L1      L2      L3      L4      L5      2
1980    39789   ------  ------  ------
1990    46410
2000    59381
2003    63531
2008    69872
2013    76410
------  ------

L2(7)=
```

VISUAL INTERPRETATION OF CORRELATION

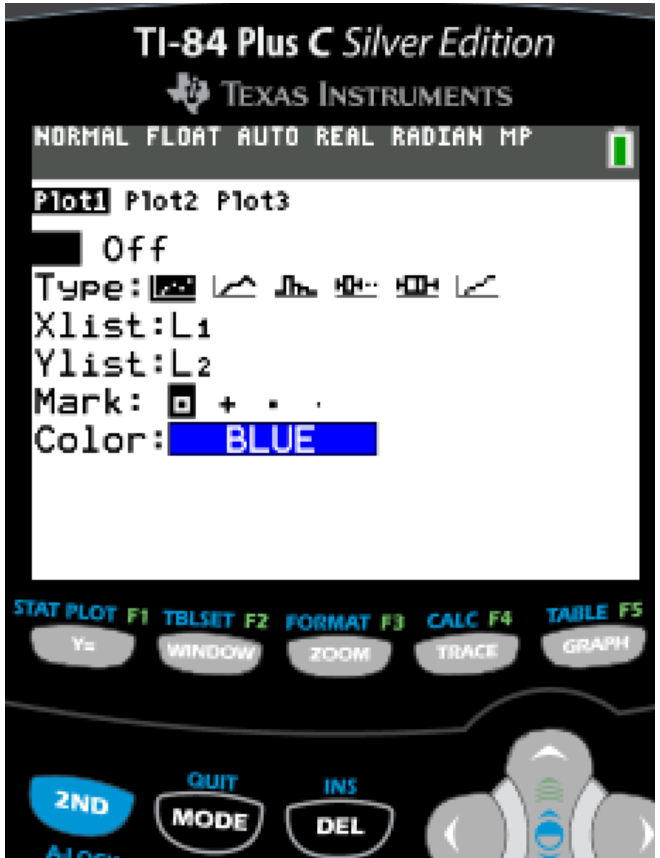
Create a Scatterplot by:

Go to STAT PLOT—Press



```
NORMAL FLOAT AUTO REAL RADIAN MP
STAT PLOTS
1:Plot1…Off
     L1    L2    ▫
2:Plot2…Off
     L1    L2    ▫
3:Plot3…Off
     L1    L2    ▫
4:PlotsOff
5:PlotsOn
```

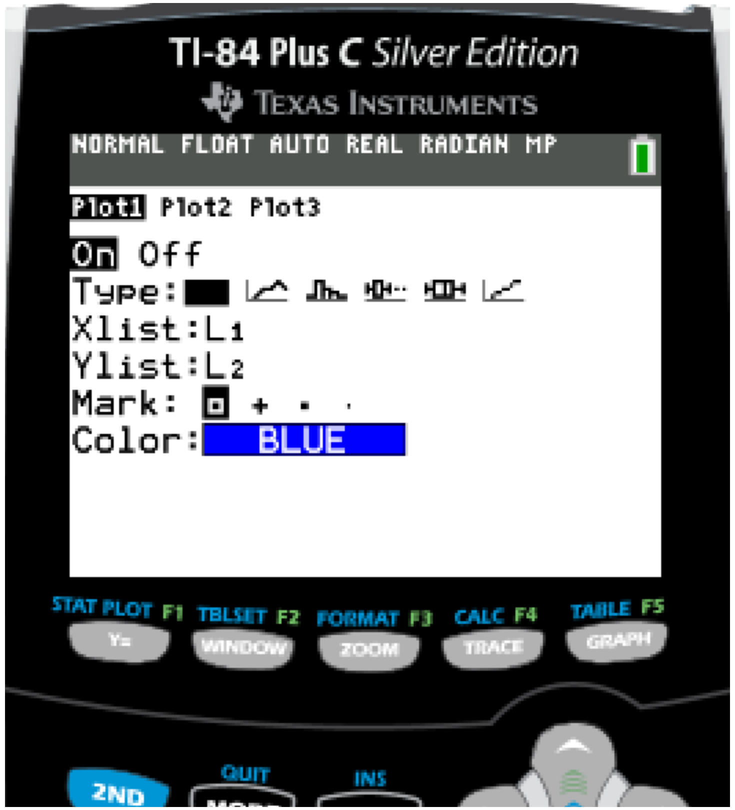
Turn on Plot 1 –choose option 1, arrow to left highlight "On"



```
NORMAL FLOAT AUTO REAL RADIAN MP
Plot1 Plot2 Plot3
     Off
Type:
Xlist:L1
Ylist:L2
Mark: ▫  +  ·  ·
Color: BLUE
```

CONTINUE SETUP OF SCATTERPLOT

Select type of graph –arrow down to “type”, arrow right 1 time

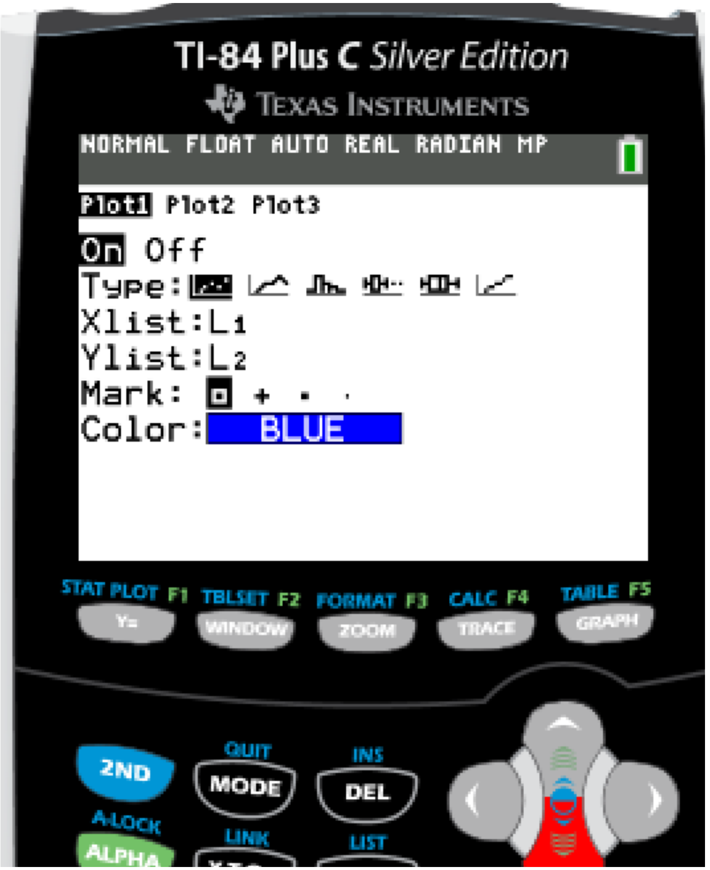


```
NORMAL FLOAT AUTO REAL RADIAN MP
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: □ + · ·
Color: BLUE
```

Indicate location of data---

Xlist: L1 (default)

Ylist = L2 (default)



```
NORMAL FLOAT AUTO REAL RADIAN MP
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: □ + · ·
Color: BLUE
```

SETUP WINDOW—SCALES FOR X- & Y- AXIS

Press TBLSET F2 WINDOW

GENERIC WINDOW
xmin= based on data given (at least lowest x value)

Xmax = based on data given (at least highest x-value)

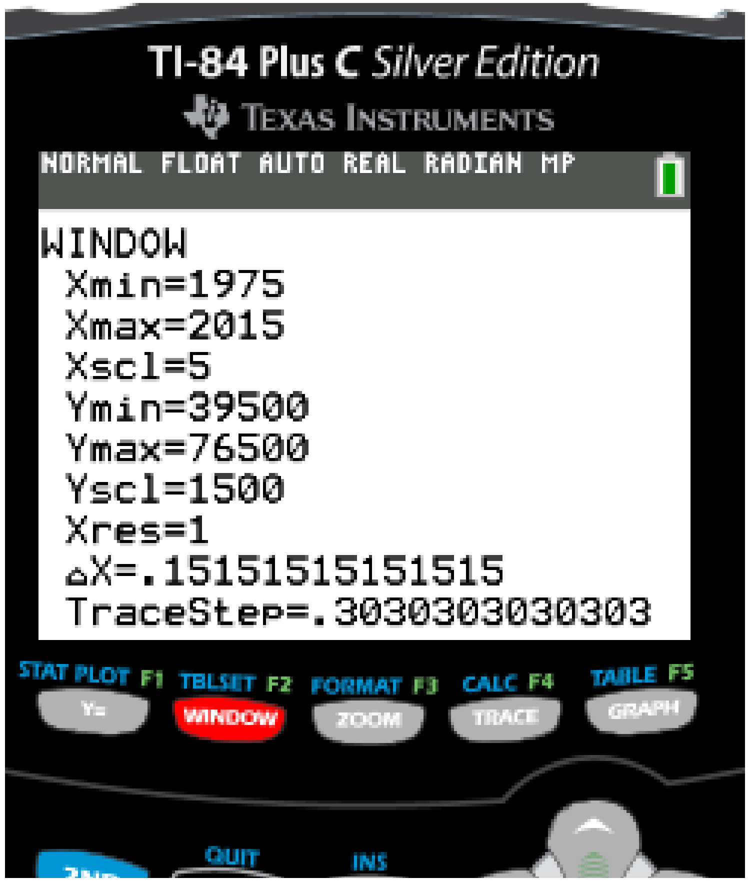
Xscl= choose a reasonable value

Ymin = based on data given (at least lowest y value)

Ymax = based on data given (at least highest y value)

Yscl = choose a reasonable value

EXAMPLE:

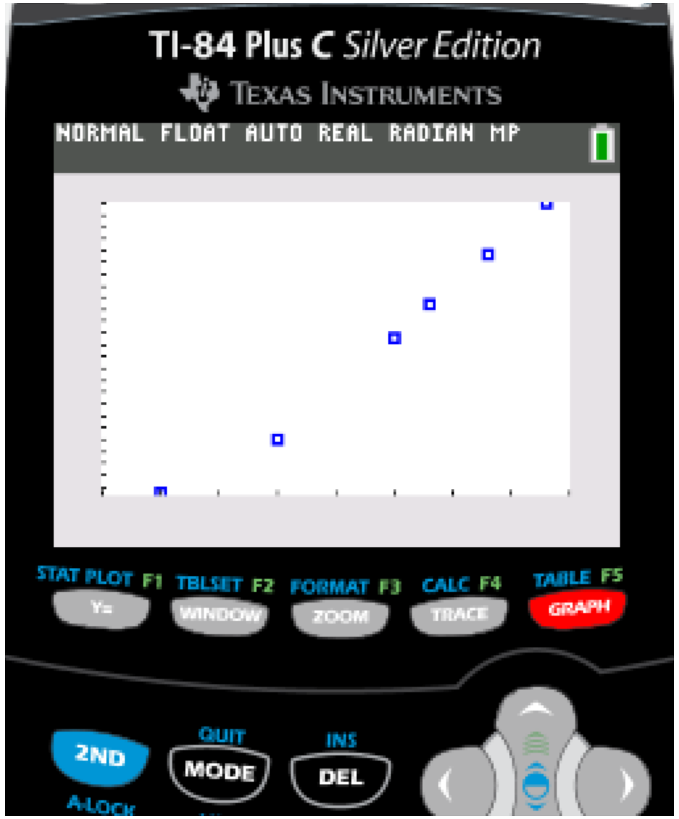


```
NORMAL FLOAT AUTO REAL RADIAN MP

WINDOW
 Xmin=1975
 Xmax=2015
 Xscl=5
 Ymin=39500
 Ymax=76500
 Yscl=1500
 Xres=1
 ΔX=.15151515151515
 TraceStep=.3030303030303
```

GRAPH

Press TABLE F5 GRAPH to see scatterplot on screen



NORMAL FLOAT AUTO REAL RADIAN MP

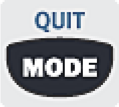
Ask yourself a few questions:

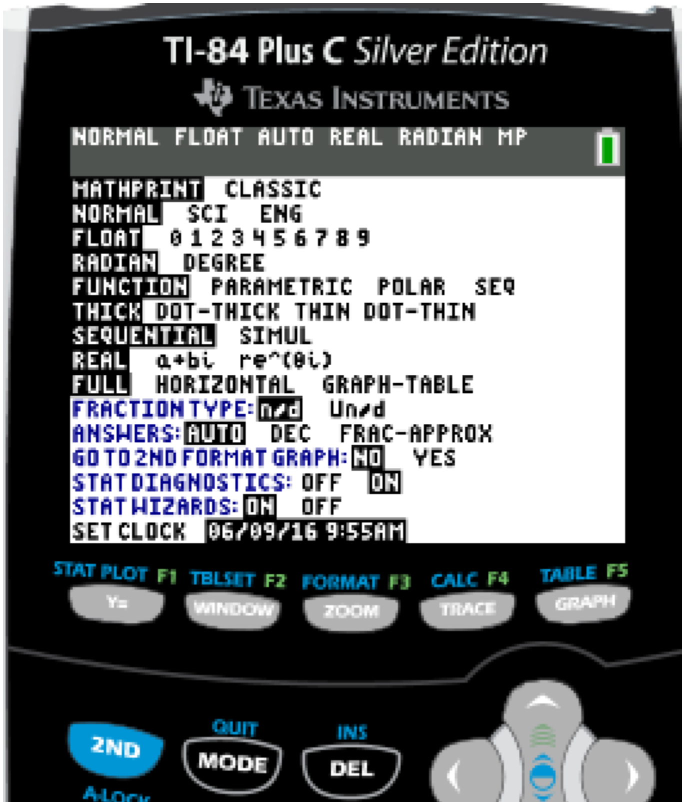
- Does there appear to be an overall pattern within the scatterplot?
- Does the overall pattern appear to be LINEAR (could you draw a line through some of the points?)?
- Does the overall pattern appear to be positive or negative?

Regardless of your responses to these questions, go to next step…

ALGEBRAIC INTERPRETATION OF CORRELATION

Turn on STAT DIAGNOSTICS

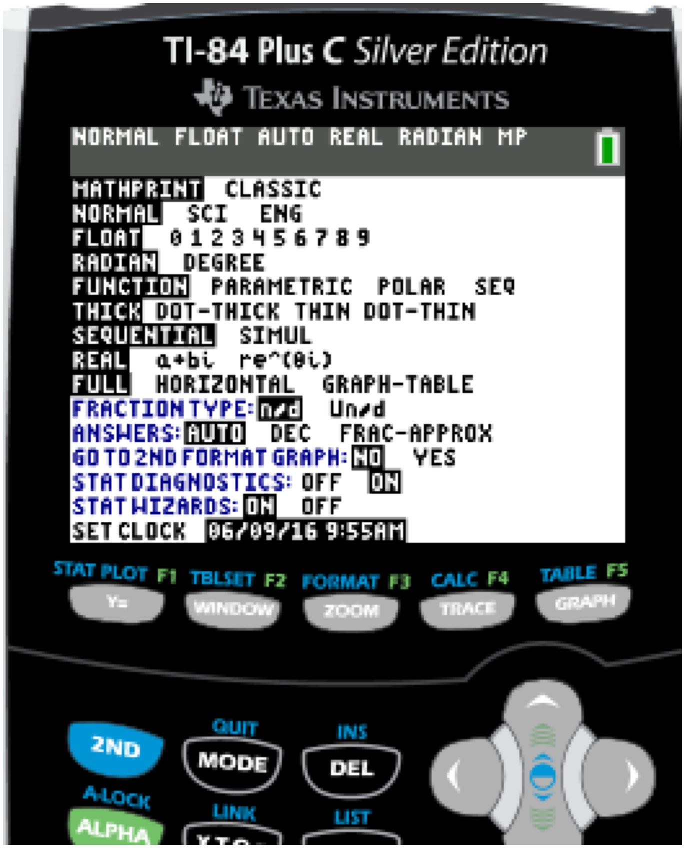




```
NORMAL FLOAT AUTO REAL RADIAN MP

MATHPRINT  CLASSIC
NORMAL  SCI  ENG
FLOAT  0 1 2 3 4 5 6 7 8 9
RADIAN  DEGREE
FUNCTION  PARAMETRIC  POLAR  SEQ
THICK DOT-THICK THIN DOT-THIN
SEQUENTIAL  SIMUL
REAL  a+bi  re^(θi)
FULL  HORIZONTAL  GRAPH-TABLE
FRACTION TYPE: n/d  Un/d
ANSWERS: AUTO  DEC  FRAC-APPROX
GO TO 2ND FORMAT GRAPH: NO  YES
STAT DIAGNOSTICS: OFF  ON
STAT WIZARDS: ON  OFF
SET CLOCK  06/09/16 9:55AM
```

Scroll down to STAT DIAGNOSTICS—highlight “ON” and enter

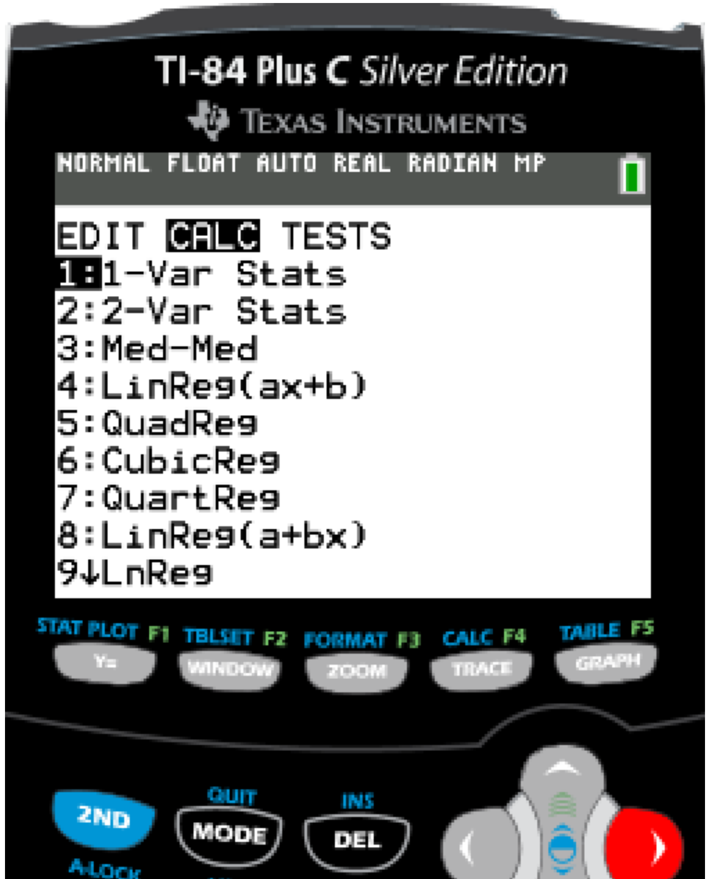


```
NORMAL FLOAT AUTO REAL RADIAN MP

MATHPRINT  CLASSIC
NORMAL  SCI  ENG
FLOAT  0 1 2 3 4 5 6 7 8 9
RADIAN  DEGREE
FUNCTION  PARAMETRIC  POLAR  SEQ
THICK DOT-THICK THIN DOT-THIN
SEQUENTIAL  SIMUL
REAL  a+bi  re^(θi)
FULL  HORIZONTAL  GRAPH-TABLE
FRACTION TYPE: n/d  Un/d
ANSWERS: AUTO  DEC  FRAC-APPROX
GO TO 2ND FORMAT GRAPH: NO  YES
STAT DIAGNOSTICS: OFF  ON
STAT WIZARDS: ON  OFF
SET CLOCK  06/09/16 9:55AM
```

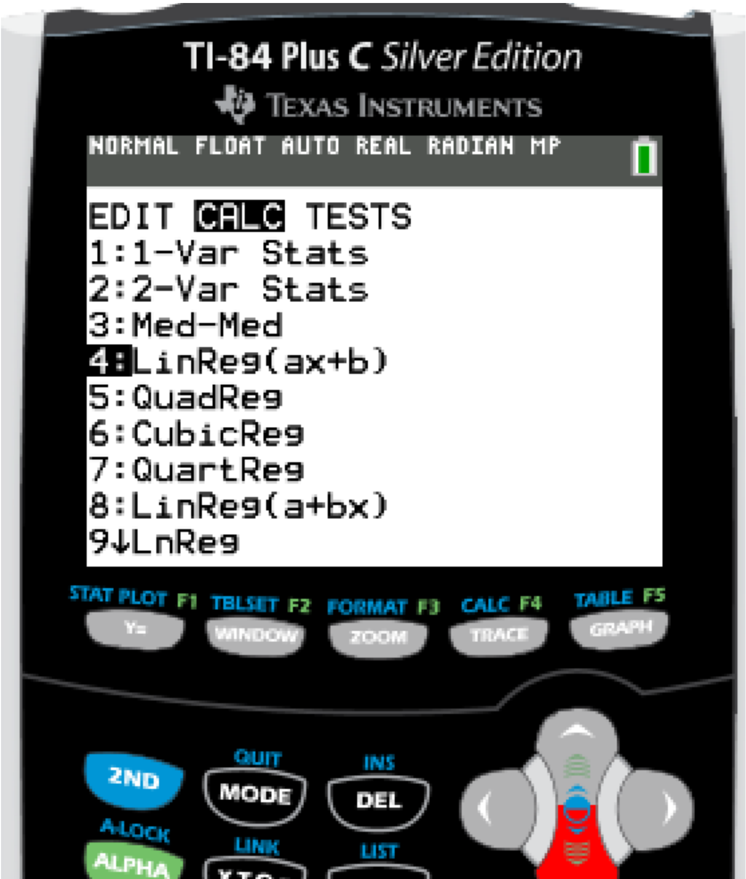
ALGEBRAIC INTERPRETATION OF CORRELATION

Press LIST STAT
Arrow right to highlight "CALC"



```
NORMAL FLOAT AUTO REAL RADIAN MP
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9↓LnReg
```

Scroll down and choose option 4:
LinReg (ax+b)



```
NORMAL FLOAT AUTO REAL RADIAN MP
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9↓LnReg
```

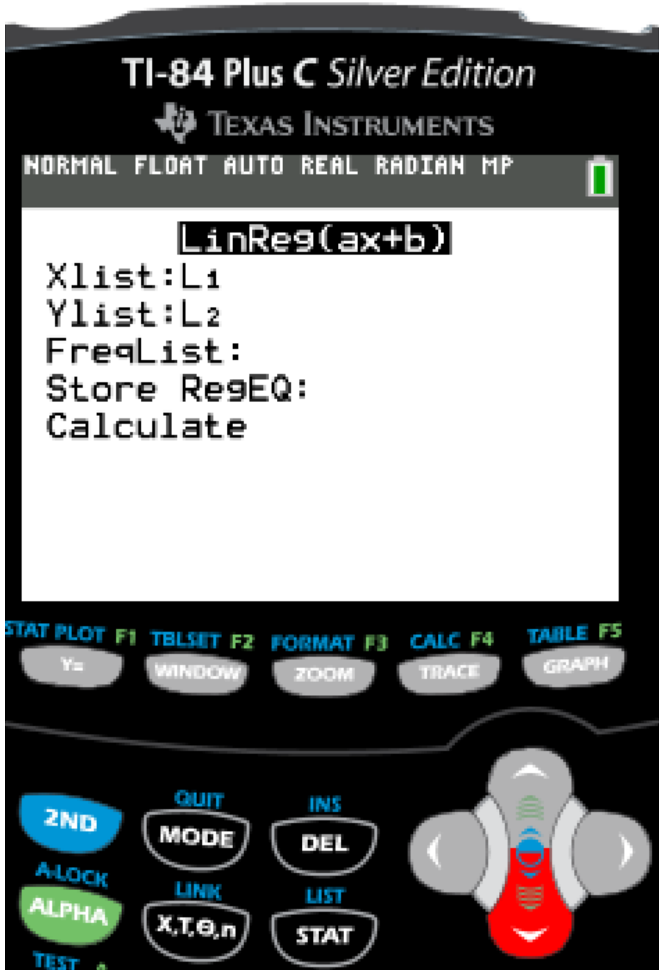
ALGEBRAIC INTERPRETATION OF CORRELATION

Indicate location of x-variable and y-variable

Enter L1 (default) –press 2ND L1 Y 1

Enter L2 (default) –press 2ND L2 Z 2

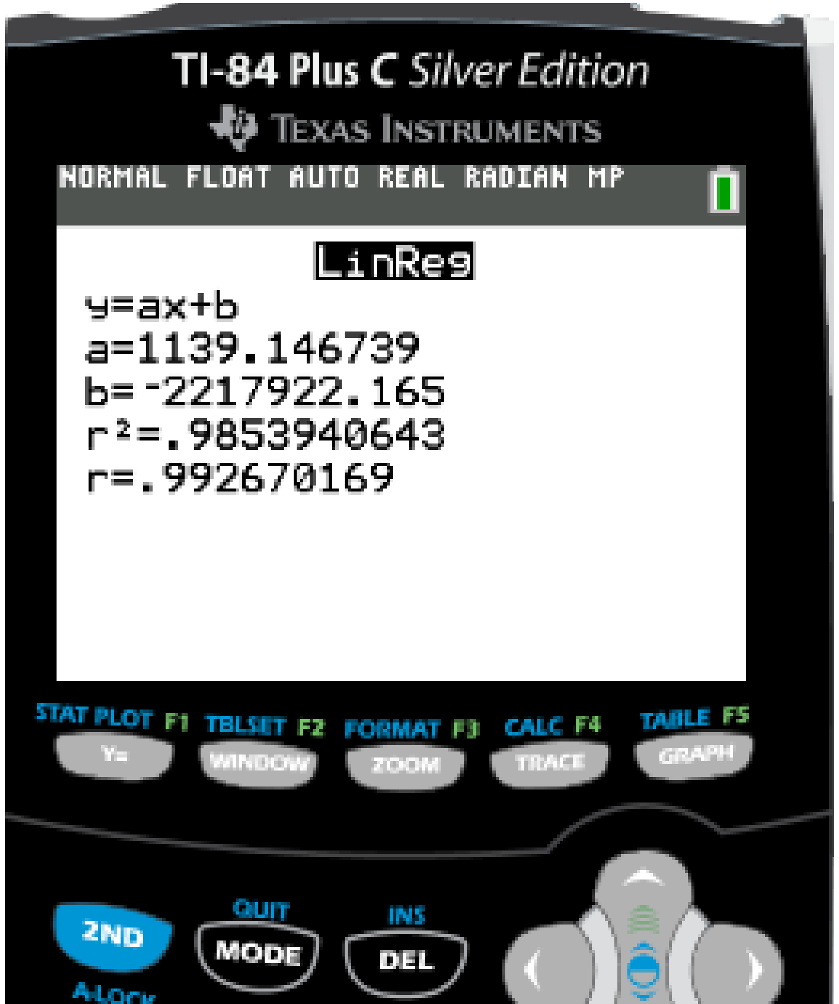
Arrow down three times to highlight "calculate"



```
NORMAL FLOAT AUTO REAL RADIAN MP
      LinReg(ax+b)
Xlist:L1
Ylist:L2
FreqList:
Store RegEQ:
Calculate
```

ALGEBRAIC INTERPRETATION OF CORRELATION

Press ENTER to see results



```
NORMAL FLOAT AUTO REAL RADIAN MP
            LinReg
 y=ax+b
 a=1139.146739
 b=-2217922.165
 r²=.9853940643
 r=.992670169
```

The sample LINEAR CORRELATION COEFFICIENT is represented by "r" on the results screen

To determine if a relationship exists and is statistically significant, compare the "r" from the calculator to the standardized "r"

GUIDELINE: If $|calculated\ r| \geq standard\ r$, then a relationship exists and is statistically significant. Therefore, proceed to next step. If the relationship is NOT statistically significant, STOP!

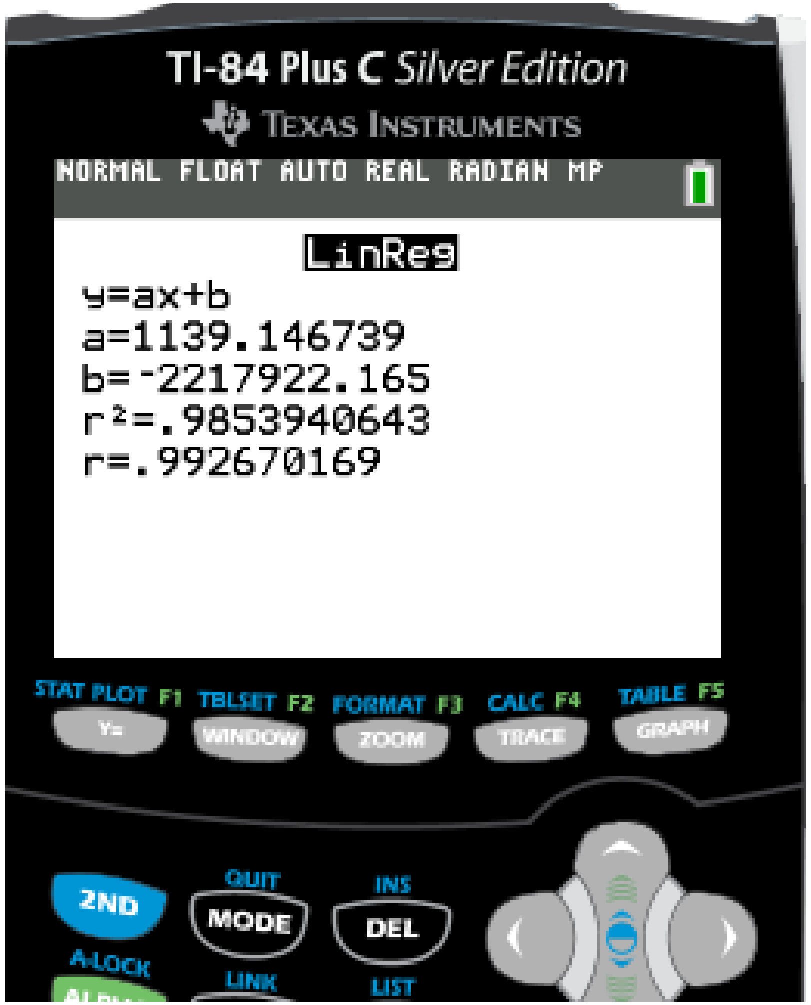
FINDING LINEAR REGRESSION EQUATION

The results window has additional information that is used to write (find) the linear regression equation that models the relationship between the variables

- a: Slope (rate of change) of the line, will be coefficient of "x"
- b: y-intercept of the line

Write in the form: $y = ax + b$

For the example, the equation would be $y = 1139.15x - 2217922.17$



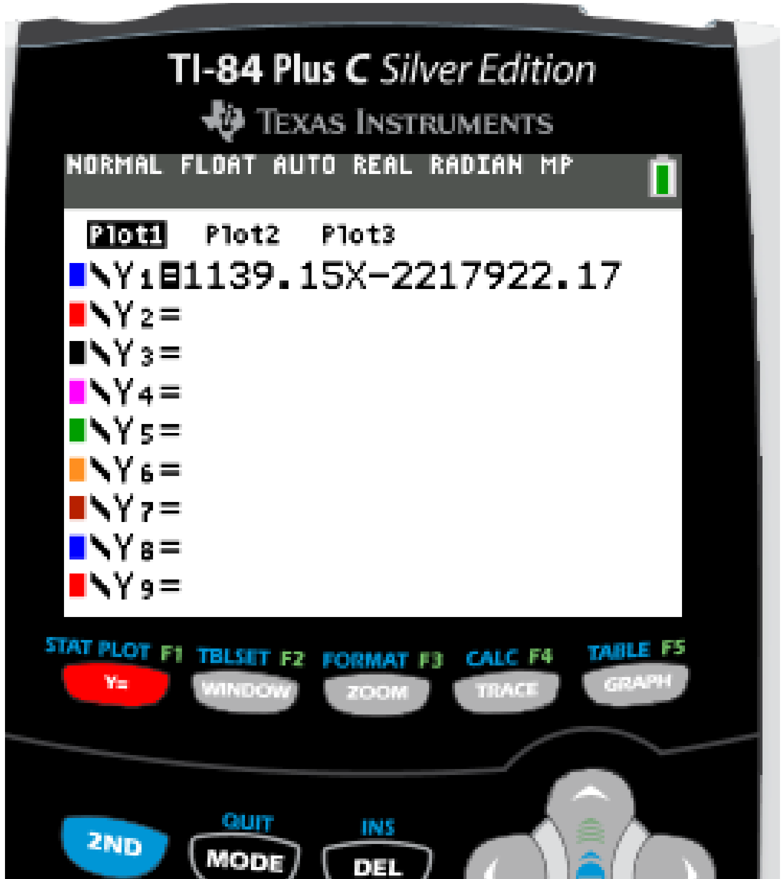
```
NORMAL FLOAT AUTO REAL RADIAN MP
LinReg
y=ax+b
a=1139.146739
b=⁻2217922.165
r²=.9853940643
r=.992670169
```

GRAPH REGRESSION LINE WITH SCATTERPLOT

Press STATPLOT F1 Y=

Enter linear equation result

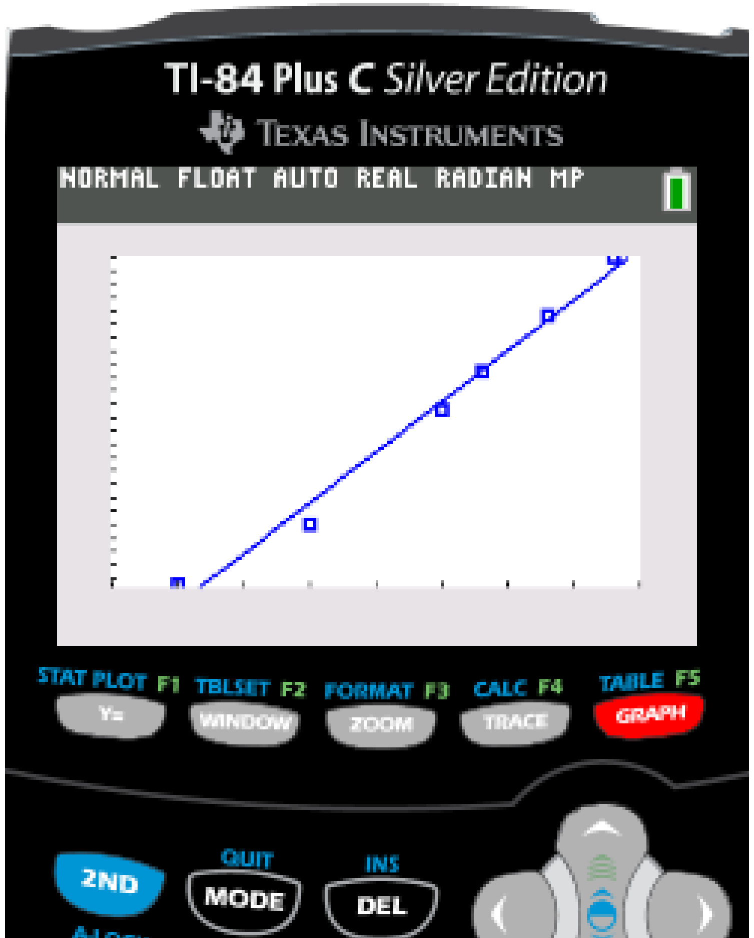
Press TABLE F5 GRAPH



NORMAL FLOAT AUTO REAL RADIAN MP

Plot1 Plot2 Plot3

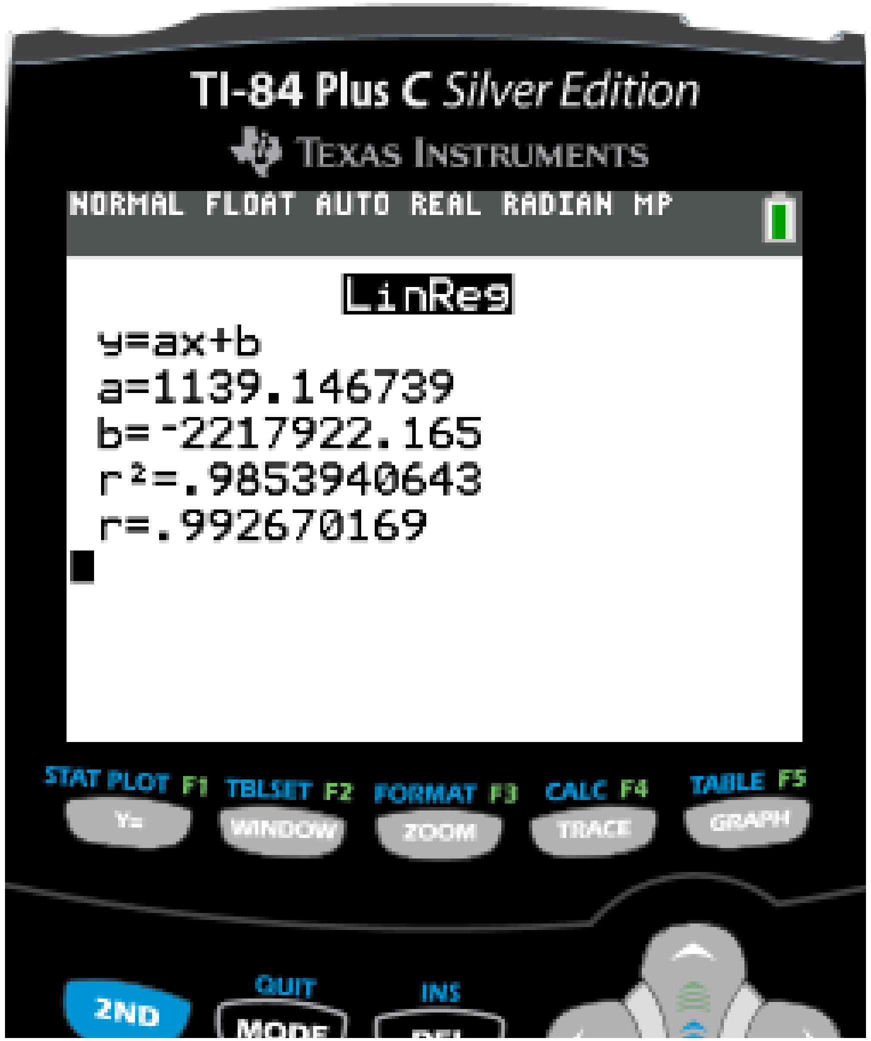
```
■\Y1=1139.15X−2217922.17
■\Y2=
■\Y3=
■\Y4=
■\Y5=
■\Y6=
■\Y7=
■\Y8=
■\Y9=
```



NORMAL FLOAT AUTO REAL RADIAN MP

COEFFICIENT OF DETERMINATION

On the results window, r^2 represents the "coefficient of determination"



```
LinReg
y=ax+b
a=1139.146739
b=-2217922.165
r²=.9853940643
r=.992670169
```

PRACTICE OF USING THE CALCULATOR

A random sample of 11statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. What is the best-fit line for this data?

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

PRACTICE OF USING THE CALCULATOR

A random sample of 11statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. What is the best-fit line for this data?

The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation

$\hat{y}$ = −173.51+4.83x

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

PREDICTION WARNING

Remember, it is always important to plot a scatter diagram first.

If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for y given x within the domain of x-values in the sample data, **but not necessarily for x-values outside that domain.**

You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the x-values in the sample data, which are between 65 and 75.

UNDERSTANDING SLOPE

The slope of the line, b, describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English.

INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

THIRD EXAM vs FINAL EXAM EXAMPLE Slope: The slope of the line is b= 4.83.

Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

THE CORRELATION COEFFICIENT "R"

Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between xand y.

The correlation coefficient, r, developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y. The correlation coefficient is calculated as

$R = \frac{n\Sigma(xy)-(\Sigma x)(\Sigma y)}{\sqrt{(n\Sigma x^2-(\Sigma x)^2)(n\Sigma y^2-(\Sigma y)^2)}}$ where n= the number of data points.

If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

(NEVER CALCULATE BY HAND)

WHAT THE VALUE OF R

- The value of r is always between –1 and +1: –1 ≤ r≤ 1.
- The size of the correlation r indicates the strength of the linear relationship between x and y. Values of r close to –1or to +1 indicate a stronger linear relationship between x and y.
- If r= 0 there is absolutely no linear relationship between x and y(no linear correlation).
- If r = 1, there is perfect positive correlation. If r = –1, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course, in the real world, this will not generally happen.

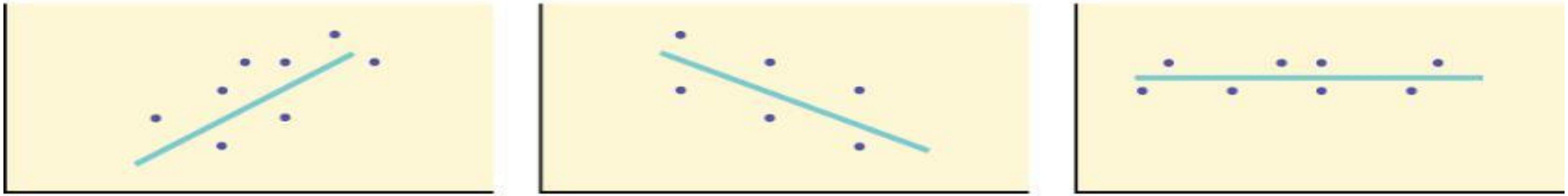
WHAT THE SIGN OF R TELLS US

- A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (positive correlation).
- A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (negative correlation).
- The sign of r is the same as the sign of the slope, b, of the best-fit line.

NOTE Strong correlation does not suggest that x causes y or y causes x.

We say “ correlation does not imply causation."

SCATTERPLOTS AND CORRELATION



(a) Positive correlation (b) Negative correlation (c) Zero correlation

(a)A scatter plot showing data with a positive correlation. $0 < r < 1$

(b)A scatter plot showing data with a negative correlation. $-1 < r < 0$

(c) A scatter plot showing data with zero correlation. $r = 0$

THE COEFFICIENT OF DETERMINATION "R^2"

The variable r^2 is called the coefficient of determination and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data:

- r^2, when expressed as a percent, represents the percent of variation in the dependent (predicted) variable y that can be explained by variation in the independent (explanatory) variable x using the regression (best-fit) line.

- $1 - r^2$,when expressed as a percentage, represents the percent of variation in y that is NOT explained by variation in x using the regression line. This can be seen as the scattering of the observed data points about the regression line.

THE COEFFICIENT OF DETERMINATION "R^2"

Consider the third exam/final exam example introduced in the previous section

- The line of best fit is: $\hat{y}= -173.51 + 4.83x$
- The correlation coefficient is $r= 0.6631$
- The coefficient of determination is $r^2 = (0.66312)^2 = 0.4397$
- Interpretation of r^2 in the context of this example:
- Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line.
- Therefore ,approximately 56% of the variation (1–0.44=0.56) in the final exam grades cannot be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.)

12.4 TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

The correlation coefficient, r, tells us about the strength and direction of the linear relationship between x and y.

However, the reliability of the linear model also depends on how many observed data points are in the sample.

We need to look at both the value of the correlation coefficient r and the sample size n, together.

We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population.

12.4 TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

The sample data are used to compute r, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient.

The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient.

The symbol for the population correlation coefficient is ρ, the Greek letter "rho."

ρ = population correlation coefficient (unknown)

r = sample correlation coefficient (known; calculated from sample data)

12.4 POPULATION CORRELATION COEFFICIENT

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient r and the sample size n.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero.

- What the conclusion means: There is a significant linear relationship between x and y. We can use the regression line to model the linear relationship between x and yin the population.

12.4 POPULATION CORRELATION COEFFICIENT

If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is not significantly different from zero."

- What the conclusion means: There is not a significant linear relationship between x and y. Therefore, we CANNOT use the regression line to model a linear relationship between x and yin the population.

REMEMBER

- If r is significant and the scatterplot shows a linear trend, the line can be used to predict the value of y for values of x that are within the domain of observed x values.
- If r is not significant OR if the scatterplot does not show a linear trend, the line should not be used for prediction.
- If r is significant and if the scatter plot shows a linear trend, the line may NOT be appropriate or reliable for prediction OUTSIDE the domain of observed x values in the data.

PERFORMING THE HYPOTHESIS TEST

- Null Hypothesis: H0: $\rho = 0$
- Alternate Hypothesis: Ha: $\rho \neq 0$

WHAT THE HYPOTHESES MEAN IN WORDS:

- Null Hypothesis H0: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship(correlation) between x and y in the population.

•Alternate Hypothesis Ha: The population correlation coefficient IS significantly DIFFERENT FROM zero. There IS A SIGNIFICANT LINEAR RELATIONSHIP (correlation) between x and y in the population.

DRAWING A CONCLUSION: METHOD 1 (USING THE P-VALUE

If the p-value is less than the significance level (α = 0.05):

- Decision: Reject the null hypothesis.
- Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from zero."

If the p-value is NOT less than the significance level (α = 0.05)

- Decision: DO NOT REJECT the null hypothesis.
- Conclusion: "There is insufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is NOT significantly different from zero."

DRAWING A CONCLUSION: METHOD 2 (USE A TABLE OF CRITICAL VALUES)

In this chapter of this textbook, we will always use a significance level of 5%, α= 0.05

The 95% Critical Values of the Sample Correlation Coefficient Table can be used to give you a good idea of whether the computed value of r is significant or not.

Compare r to the appropriate critical value in the table.

If r is not between the positive and negative critical values, then the correlation coefficient is significant.

If r is significant, then you may want to use the line for prediction.

PRACTICE OF USING THE CALCULATOR

A random sample of 11statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200. Can we use the line to predict the final exam score (predicted y value)?

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

CAN WE USE THE REGRESSION LINE TO PREDICT FINAL EXAM SCORES

H0: ρ = 0

Ha: ρ ≠ 0 α = 0.05

- Use the "95% Critical Value" table for r with df=n– 2 = 11 – 2 = 9.
- The critical values are –0.602 and +0.602
- Since 0.6631 > 0.602, r is significant.
- Decision: Reject the null hypothesis.
- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between the third exam score (x) and the final exam score (y) because the correlation coefficient is significantly different from zero.

Because r is significant and the scatterplot shows a linear trend, the regression line can be used to predict final exam scores.

ASSUMPTIONS IN TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

Testing the significance of the correlation coefficient requires that certain assumptions about the data are satisfied. The premise of this test is that the data are a sample of observed points taken from a larger population.

We have not examined the entire population because it is not possible or feasible to do so. We are examining the sample to draw a conclusion about whether the linear relationship that we see between x and y in the sample data provides strong enough evidence so that we can conclude that there is a linear relationship between x and y in the population.

ASSUMPTIONS IN TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

The regression line equation that we calculate from the sample data gives the best-fit line for our particular sample. We want to use this best-fit line for the sample as an estimate of the best-fit line for the population. Examining the scatterplot and testing the significance of the correlation coefficient helps us determine if it is appropriate to do this.

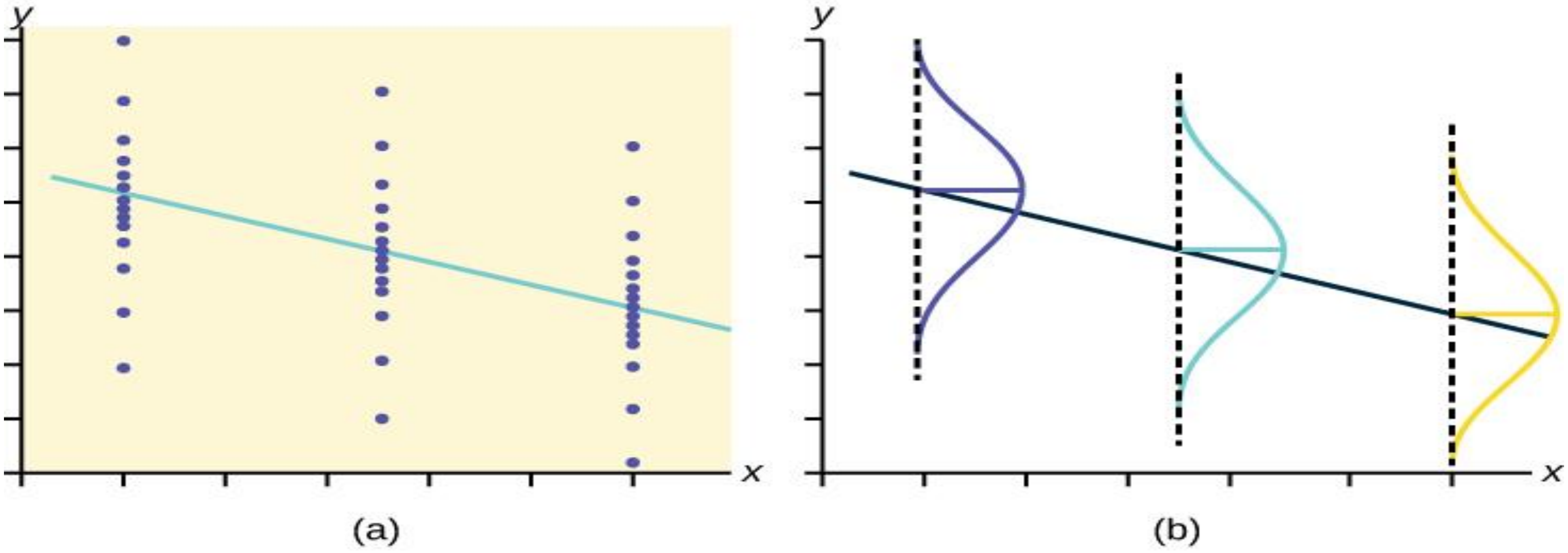
The assumptions underlying the test of significance are:

- There is a linear relationship in the population that models the average value of y for varying values of x. In other words, the expected value of y for each particular value lies on a straight line in the population. (We do not know the equation for the line for the population. Our regression line from the sample is our best estimate of this line in the population.)

ASSUMPTIONS IN TESTING THE SIGNIFICANCE OF THE CORRELATION COEFFICIENT

- They values for any particular x value are normally distributed about the line. This implies that there are more y values scattered closer to the line than are scattered farther away. Assumption (1) implies that these normal distributions are centered on the line: the means of these normal distributions of y values lie on the line.
- The standard deviations of the population y values about the line are equal for each value of x. In other words, each of these normal distributions of y values has the same shape and spread about the line.
- The residual errors are mutually independent (no pattern).
- The data are produced from a well-designed, random sample or randomized experiment.

EXAMPLE OF NORMAL DISTRIBUTED DATA



The *y* values for each *x* value are normally distributed about the line with the same standard deviation. For each *x* value, the mean of the *y* values lies on the regression line. More *y* values lie near the line than are scattered further away from the line.

12.5 PREDICTION

PREDICTION

Recall the third exam/final exam example. We examined the scatterplot and showed that the correlation coefficient is significant. We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores(x-values) range from 65 to 75. Since73 is between the x-values 65 and 75, substitute x=73 into the equation.

Then: $\hat{y} = -173.51+4.83(73)=179.08$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

PREDICTION

Recall the third exam/final exam example.

y-hat = −173.51+4.83(x)

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

PREDICTION

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

145.27

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

The x values in the data are between 65 and 75. Ninety is outside of the domain of the observed x values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for x and calculate a corresponding y value, the y value that you get will not be reliable.) To understand really how unreliable the prediction can be outside of the observed x values observed in the data, make the substitution x= 90 into the equation.

$\hat{y} = -173.51+4.83(90) = 261.19$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

INTERPOLATION AND EXTRAPOLATION

The process of predicting inside of the observed x values observed in the data is called **interpolation.**

The process of predicting outside of the observed x values observed in the data is called **extrapolation.**

12.6 OUTLIERS

OUTLIERS

In some data sets, there are values (observed data points) called outliers. Outliers are observed data points that are far from the least squares line. They have large "errors", where the "error" or residual is the vertical distance from the line to the point.

Outliers need to be examined closely. Sometimes, for some reason or another, they should not be included in the analysis of the data. It is possible that an outlier is a result of erroneous data.

Other times, an outlier may hold valuable information about the population under study and should remain included in the data. The key is to examine carefully what causes a data point to be an outlier.

OUTLIERS

Besides outliers, a sample may contain one or a few points that are called **influential points**. Influential points are observed data points that are far from the other observed data points in the horizontal direction. These points may have a big effect on the slope of the regression line. To begin to identify an influential point, you can remove it from the data set and see if the slope of the regression line is changed significantly.

Computers and many calculators can be used to identify outliers from the data. Computer output for regression analysis will often identify both outliers and influential points so that you can examine them.

IDENTIFYING OUTLIERS

We could guess at outliers by looking at a graph of the scatterplot and best fit-line. However, we would like some guideline as to how far away a point needs to be in order to be considered an outlier. **As a rough rule of thumb, we can flag any point that is located further than two standard deviations above or below the best-fit line as an outlier. The standard deviation used is the standard deviation of the residuals or errors.**

We can do this visually in the scatterplot by drawing an extra pair of lines that are two standard deviations above and below the best-fit line. Any data points that are outside this extra pair of lines are flagged as potential outliers. Or we can do this numerically by calculating each residual and comparing it to twice the standard deviation. On the TI-83, 83+, or 84+, the graphical approach is easier. The graphical procedure is shown first, followed by the numerical calculations. You would generally need to use only one of these methods.

GRAPHICALLY IDENTIFYING OUTLIERS

If we were to measure the vertical distance from any data point to the corresponding point on the line of best fit and that distance were equal to 2s or more, then we would consider the data point to be "too far“ from the line of best fit. We need to find and graph the lines that are two standard deviations below and above the regression line. Any points that are outside these two lines are outliers. We will call these lines Y2 and Y3:

Remember Y1 or $\hat{y}$ is the line of best fit. Using the third exam and final exam data, the line of best fit was

$\hat{y} = -173.51+4.83(x)$

GRAPHICALLY IDENTIFYING OUTLIERS

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find s= **16.412**.

Since the data is already in the calculator, Go to STAT, TESTS, down to **LinRegTTest**. Scroll down until you see **s.**

$\hat{y}$ = −173.51+4.83(x)

Line Y2 = −173.51+4.83(x) – 2(16.4) (2 standard deviations below $\hat{y}$)

Line Y3 = −173.51+4.83(x) + 2(16.4) (2 standard deviations above $\hat{y}$)

GRAPHICALLY IDENTIFYING OUTLIERS

1. Make a scatterplot using $\hat{y}$ = −173.51+4.83(x)

2. Enter the two lines in the y=

Line Y2 = −173.51+4.83(x) – 2(16.4)

Line Y3 = −173.51+4.83(x) + 2(16.4)

3. Press ZOOM and then 9.   9 is ZoomStat

4. You will find that the only data point that is not between lines Y2 and Y3 is the point x = 65, y = 175. On the calculator screen it is just barely outside these lines. The outlier is the student who had a grade of 65 on the third exam and 175 on the final exam; this point is further than two standard deviations away from the best-fit line.

NUMERICAL IDENTIFICATION OF OUTLIERS

To find the numerical identification of outliers, the first step is to make a table to put the values in. It might be easy to complete this table in excel. Again I am using the third exam and final exam data.

- The first column will be your x-values.
- The second column will be your y-values.
- The third column will be the value you get when you pull x in the line of best fit ($\hat{y}$)
- The fourth column will be y (second column) - $\hat{y}$ (third column). We call these the residuals.

NUMERICAL IDENTIFICATION OF OUTLIERS

x	y	$\hat{y}$	y - $\hat{y}$ (residuals)
65	175	-175.5 + 4.83(65) = 140	175 – 140 = 35
67	133	-175.5 + 4.83(67) = 150	133 – 150 = -17
71	185	-175.5 + 4.83(71) = 169	185 – 169 = 16
71	163	-175.5 + 4.83(71) = 169	163 – 169 = -6
66	126	-175.5 + 4.83(66) = 145	126 – 145 = -19
75	198	-175.5 + 4.83(75) = 189	198 – 189 = 9
67	153	-175.5 + 4.83(67) = 150	153 – 150 = 3
70	163	-175.5 + 4.83(70) = 164	163 – 164 = -1
71	159	-175.5 + 4.83(71) = 169	159 – 169 = -10
69	151	-175.5 + 4.83(69) = 160	151 – 160 = -9
69	159	-175.5 + 4.83(69) = 160	159 – 160 = -1

NUMERICAL IDENTIFICATION OF OUTLIERS

As we did with the equation of the regression line and the correlation coefficient, we will use technology to calculate this standard deviation for us. Using the **LinRegTTest** with this data, scroll down through the output screens to find s= **16.412**.

Since the data is already in the calculator, Go to STAT, TESTS, down to **LinRegTTest.** Scroll down until you see **s.**

We are looking for all data points for which the residual is

- greater than 2s = 2(16.4) = 32.8 or
- less than -2s = -2(16.4) = -32.8

The point (65,175) has a residual of 35 so it is outside the range.

It is a potential outlier.

REMOVING OUTLIERS

Numerically and graphically, we have identified the point (65, 175) as an outlier.

We should re-examine the data for this point to see if there are any problems with the data.

- If there is an error, we should fix the error if possible, or delete the data.
- If the data is correct, we would leave it in the data set.

For this problem, we will suppose that we examined the data and found that this outlier data was an error. Therefore we will continue on and delete the outlier, so that we can explore how it affects the results, as a learning experience.

REMOVING OUTLIERS

Compute a new best-fit line and correlation coefficient using the ten remaining points:

1. On the TI-83, TI-83+, TI-84+ calculators, delete the outlier from L1 and L2.
2. Using the LinReg(ax+b) or LinRegTTest, the new line of best fit and the correlation coefficient are:

$$\hat{y} = -355.19 + 7.39x \text{ and } r = 0.9121$$

The new line with r=0.9121 is a stronger correlation than the original (r=0.6631) because r=0.9121is closer to one. This means that the new line is a better fit to the ten remaining data values. The line can better predict the final exam score given the third exam score.