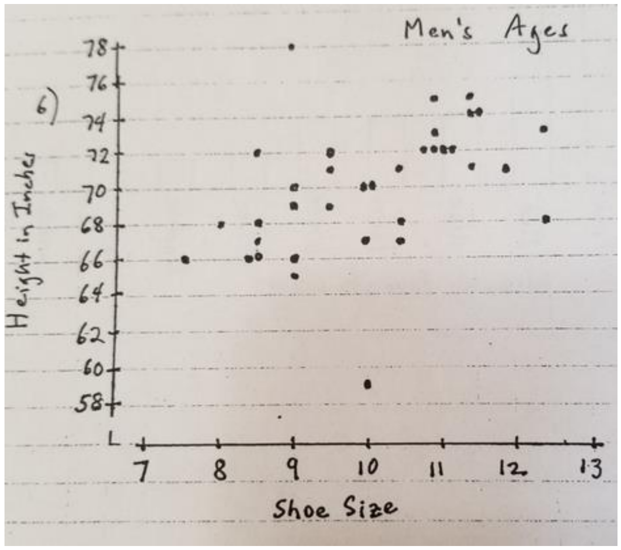
Online Assignment #7: Correlation and Regression

Remember in Assignment #1 you made a scatterplot of men's shoe size and height? This is what it looked like:

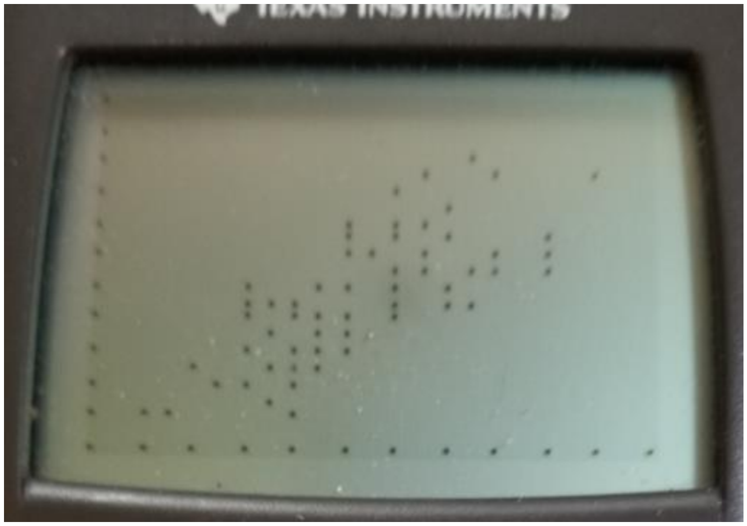


In general, men who wear larger shoes are taller than those who wear smaller shoes. We call this **correlation**, co-relation, a connection between two variables. Sometimes it's a **positive** correlation – we see that as shoe size increases, so does height. Of course there are outliers, like the person who wears a size 10 shoe and is 59 inches tall, and the person who wears a size 9 and is 78 inches tall.

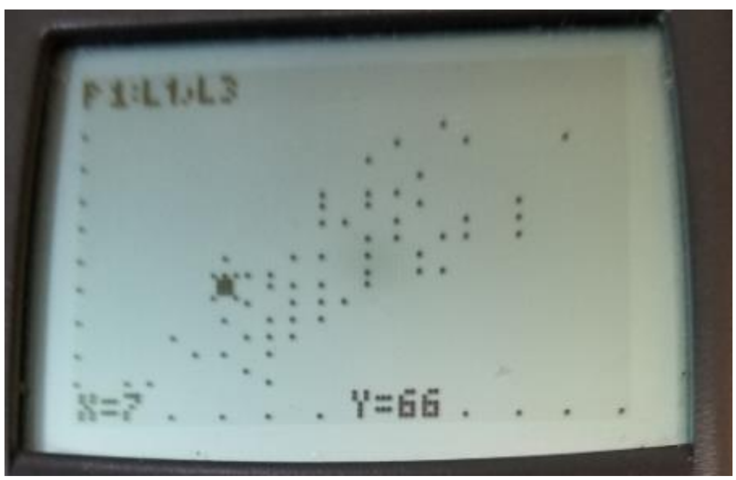
We want to do more than just say that shoe size and height are related: we want to have a way to predict (guess) a man's height from his shoe size. To do that, we use **linear regression**, in which we make a straight line that goes the closest to the most dots.

Let's use the whole Class Data Base here, and let's look at shoe size and height for the entire group, using shoe size as the *x*-, or independent, or **predictor** variable, and height as the *y*-, or dependent, or **response** variable.

Here's the scatterplot on the calculator:



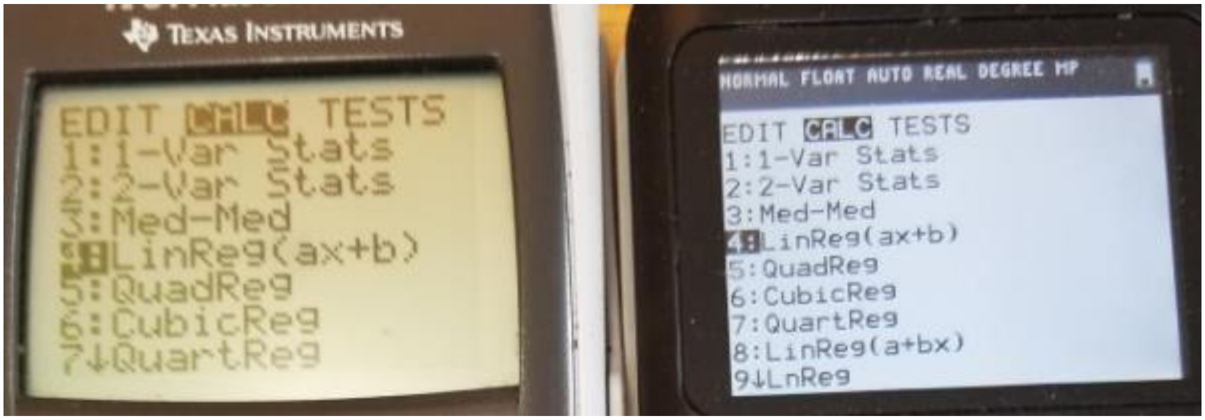
The calculator doesn't show duplicates in either of the ways described in Lecture #3. It does, however, display the *x*- and *y*-values of each point, using the Trace function:



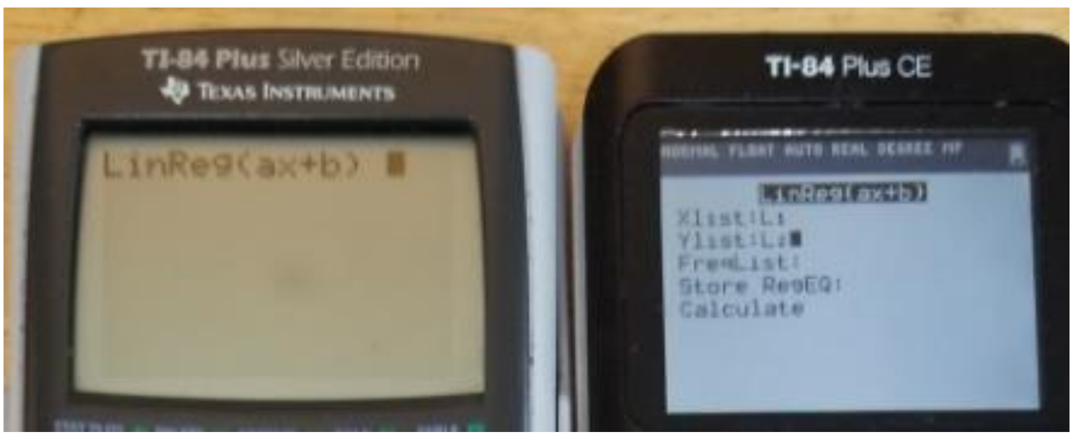
I'm not going to explain how to make the scatterplot on your calculator. If you want to know, go to YouTube.

What I will explain is how to find the equation we would use if we wanted to predict a person's height from their shoe size. We call it the **least-squares, best-fit regression line**, and you can read about what its name means on p. 194 in the text.

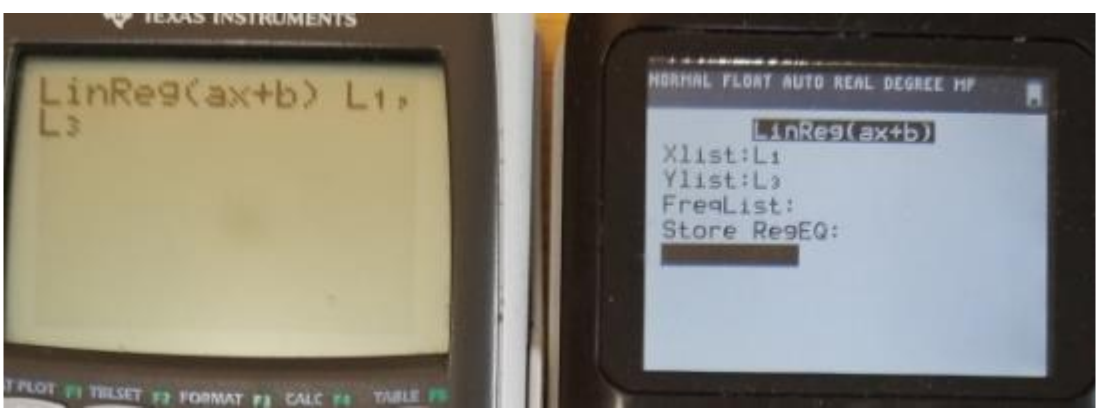
First, go to the Stat Calc menu and choose Option 4: LinReg(ax+b):



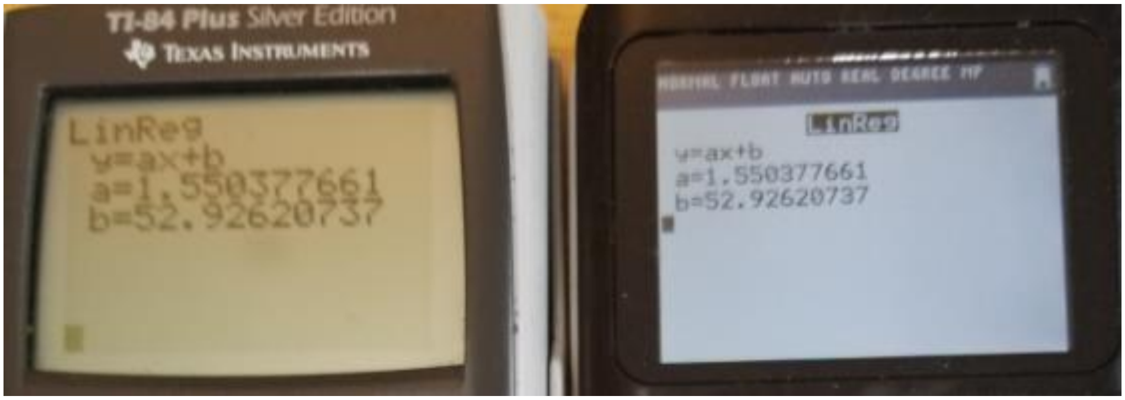
Your screen will then look like this:



With the old system, put in the *x*-list name and the *y*-list name separated by a comma. In the new system, be sure to leave FreqList empty, and you don't have to worry about Store RegEQ:



The result might look like this:



(If your calculator has some more information on it, don't worry. We will all catch up with you shortly.)

So *a* is the slope of the line, and *b* is the *y*-intercept. If you've forgotten about equations of straight lines, YouTube's the place for you. Anyway, the equation we would use to predict height from shoe size, with *a* and *b* rounded to the nearest thousandth, is

$$y = 1.550x + 52.926$$

Here's how you use it. Let's say I want to predict the height of a person who wears a size 9½ shoe. I put 9.5 in for *x* (what we call **evaluating the expression for *x* = 9.5**):

$$y = 1.550 \cdot 9.5 + 52.926 = 67.651$$

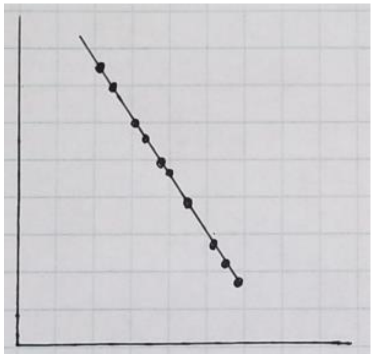
or 67.7 inches to the nearest tenth.

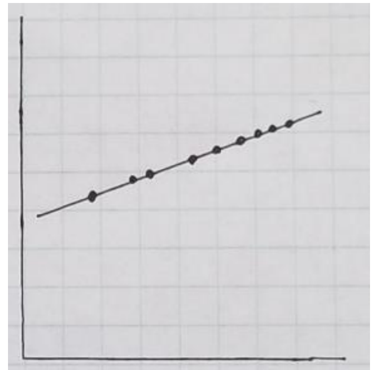
Three important points about this:

1) We're not saying that the size of your shoes **causes** your height to be what it is. That would be ridiculous. We're saying that shoe size and height are both products of some basic size gene

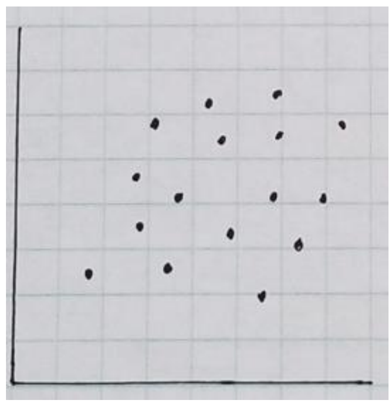
2) You can't use this equation to make predictions unless you can show that there is a true connection between the variables and that the pattern of dots isn't some random outcome. To do that, we find a number, called the **correlation coefficient**, ***r***, which tells us how closely the points hug the line. The way it's computed, it has to be between -1 and $+1$.

It's $\mathbf{-1}$ if the points **all** lie on a straight line, and that line has a **negative** slope, and it's $\mathbf{+1}$ if the points **all** lie on a straight line, and that line has a **positive** slope:

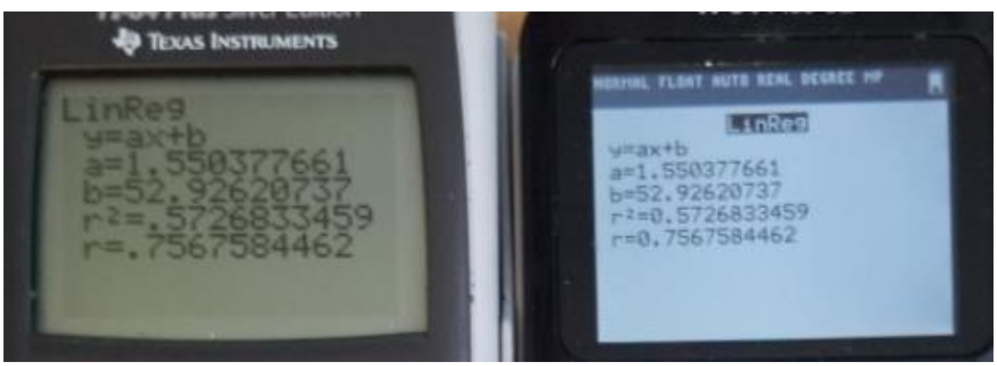




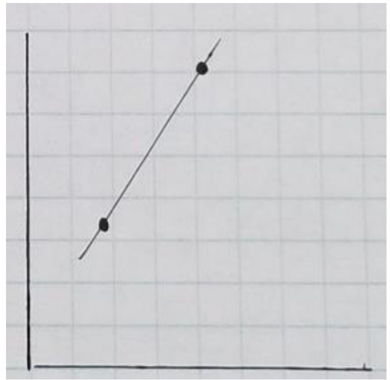
And it's close to 0 if the points have no particular pattern:



Your calculator may already be reporting the value of *r*, but if your calculator display looks like the one on p. 4, you can get your calculator to do so by following the instructions on p. 203 of the text. After that, your screen should look like this:

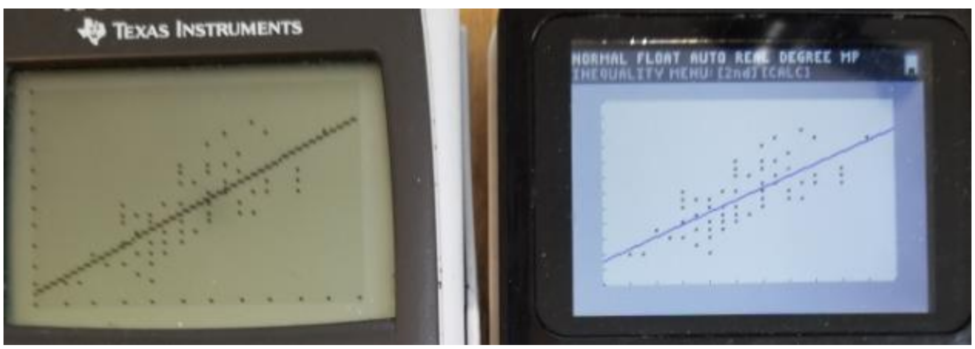


So $r \approx 0.757$. Does this mean the points are close enough to the line to use the equation to predict height from shoe size? That depends on how many points are on the scatterplot. If there are only two, of course they lie on a straight line. As Euclid put it (in Greek, though), two points determine a line:

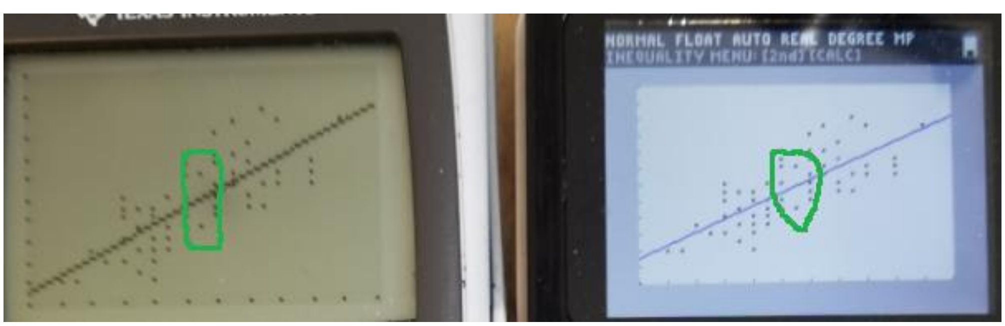


To determine how close to -1 or $+1$ r needs to be, we need to know how many points there are in the scatterplot, and this brings up the concept of **degrees of freedom**, which I'm going to gloss over. Let's just say that if $r < -0.350$ or $r > 0.350$ we can use the equation to predict. That will be the cutoff.

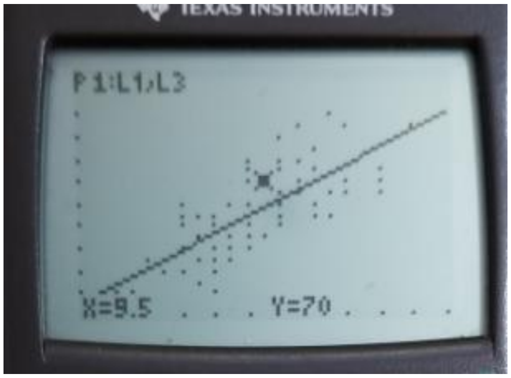
Look at the scatterplot with the line included:

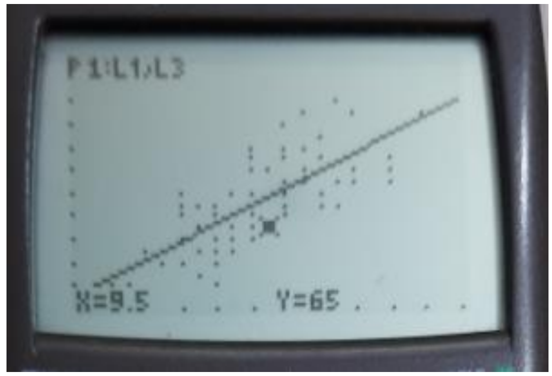


3) We calculated, using the regression equation, that a person who wears a size 9½ shoe should be about 67.7 inches tall. From the photo below, you can see that there are two people in the data base who wear a size 9½ shoe.



The one above the line is Person #9, who is 70 inches tall, and the one below the line is Person #70, who is 65 inches tall:





Using the **actual** height of the people who wear size 9½ shoes and the **predicted** height of the people who wear size 9½ shoes, we have a concept called the **residual**, which is the **actual height minus the predicted height**.

For Person # 9, this would be $70 - 67.7 = 2.3$ inches. Person #9 is 2.3 inches **taller** than you'd predict for someone who wears size 9½ shoes.

For Person #70, it would be $65 - 67.7 = -2.7$ inches. Person #70 is 2.7 inches **shorter** than you'd predict for someone who wears size 9½ shoes.

Don't worry: we won't be chopping Person #9 down to size or stretching Person #70. The residual just tells how the person's actual height compares to the predicted. If a person is taller than predicted, their residual is positive; if shorter than predicted, their residual is negative.

This is how your answer would look if the instructions were:

a) Using shoe size as the predictor variable and height as the response variable, state the equation of the regression line and the correlation coefficient to the nearest thousandth.

b) Predict to the nearest tenth the height of a person who wears size 9½ shoes.

c) Find the residuals for Persons #9 and #70 to the nearest tenth.

a) $y = 1.550x + 52.926$; $r \approx 0.757$

b) 67.7 in.

c) Person #9: 2.3 inches
Person #70: -2.7 inches

Assignment

1) a) Using height as the predictor variable and shoe size as the response variable, state the equation of the regression line and the correlation coefficient to the nearest thousandth.

b) Predict to the nearest tenth the shoe size of a person who is 67 inches tall.

c) Find the residuals for Persons #71 and #29 to the nearest tenth.

2) Using your project numbers, with number of pages as the predictor variable and thickness as the response variable,

a) State the equation of the regression line and the correlation coefficient to the nearest thousandth.

b) List two reasons why *r*, the correlation coefficient, isn't equal to 1. In other words, why don't all books that have the same number of pages have the same thickness?

c) Pick two numbers of pages from your list and use your equation in 2)a) to predict the thickness for each number of pages, to the nearest tenth of a millimeter.

d) For your choices in c) calculate the residuals to the nearest tenth of a millimeter (actual thickness minus predicted thickness).

e) Make a scatterplot of your data, using number of pages for the x-variable and thickness for the y-variable. (If you don't have graph paper, just do your best.) Graph your equation from 2)a) on the scatterplot. First let x = 0, and that will give you one point. Then to get a second point, pick another number of pages and calculate its predicted thickness. Plot the two points and draw the line through them.