Chapter 5 – Scatterplots, Correlation, and Regression

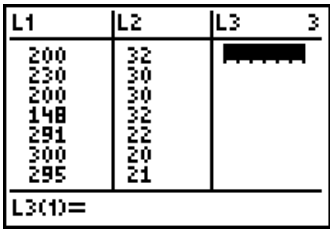
Are two numeric variables related? If so, how? Scatterplots and regression will answer these questions. Correlation describes the direction and strength of linear relationships. Linear regression further describes these relationships.

Here are advertised horsepower ratings and expected highway gas mileage for several model year 2007 vehicles. These are the data for problem 35 in Chapter 7 of the text.

Audi A4	200 hp	32 mpg	Honda Accord	166	34
BMW 328	230	30	Hyundai Elantra	138	36
Buick LaCrosse	200	30	Lexus IS 350	306	28
Chevy Cobalt	148	32	Lincoln Navigator	300	18
Chevy Trailblazer	291	22	Mazda Tribute	212	25
Ford Expedition	300	20	Toyota Camry	158	34
GMC Yukon	295	21	VW Beetle	150	30
Honda Civic	140	40			

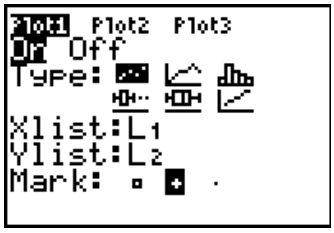
How is horsepower related to gas mileage? The first step in examining relationships is through a scatterplot.

SCATTERPLOTS

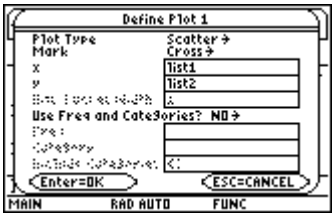


Here are the first few values for the horsepower ratings (in L1) and the gas mileages (in L2). It is important to enter these very carefully as they have been entered in the table, since the values represent a data pair for each vehicle type.

Our supposition is that larger engines will get less gas mileage, so we will use the horsepower ratings as the predictor (*X*) variable and the gas mileage as the response (*Y*) variable.

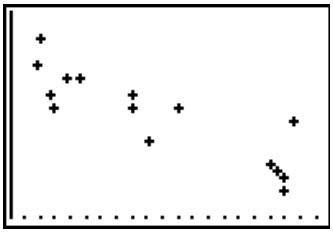


To define the scatterplot, press 2nd Y= (STAT PLOT). Select Plot1 by pressing ENTER. Scatterplots are the first plot type. Move the cursor to highlight that plot, and press ENTER to move the highlight. Press the down arrow (▼) and enter the list where the predictor (*X*) variable is (here, L1, so 2nd 1). Press the down arrow and enter the list containing the response (*Y*) variable (here, L2, so 2nd 2). Press ▼ to select the type of mark for each data point (*X*, *Y*) pair. The author of this manual recommends either the square or cross; the single pixel tends to be too hard to see. When finished, the plot definition screen should look like the one at right.



To define the scatterplot on a TI-89, press F2 (Plots) then select choice 1:Plot Setup. Make sure unnecessary plots are cleared out by moving the cursor to highlight the plot and pressing F3 to clear the plot or F4 to uncheck it. Press F1 to start defining Plot 1. Press the right arrow to select the plot type. Press ENTER to select 1:Scatter. Press the down arrow (▼) and select your choice of Mark for the data points by using the right arrow to expand the menu, then the down arrow to make your selection. The author recommends against the last choice 5:Dot as the single pixel tends to be too hard to see. Press the down arrow and select the list with the predictor (*X*) values (2nd − gets the [VAR-LINK]screen; move the cursor to the desired list and press ENTER). Press the down arrow and enter the list containing the response (*Y*). When finished, the plot definition screen should look like the one at right.

Press ZOOM 9 (F5 on an 89) to view the plot. Here we see a generally decreasing pattern from left to right, supporting our initial idea. The pattern is generally linear; however, some points at the bottom right may be unusual; we'll examine those later.

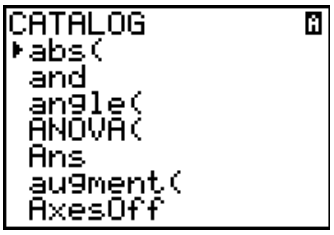


REGRESSION SETUP (TI-83/84 ONLY)

We want to examine the nature of this relationship further; before we do, we need to set up the calculator to display the values of the correlation coefficient (r) and the coefficient of determination (r^2). This only needs to be done if you are using a TI-83/84. The TI-89 series will always display these quantities.

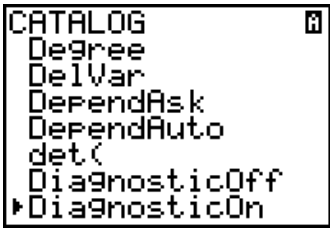
This procedure *normally* needs to be done only *once*; however, changing batteries slowly will reset memory and it may have to be done again.

Press 2nd 0 (Catalog). This accesses the list of all the functions the calculator knows about. Notice that the cursor is at the beginning of the catalog. We need to get down to a command that begins with a D, so press x^{-1} (The equivalent of alpha D. The cursor is already in Alphabetic mode, indicated by the A at the upper right on the screen).

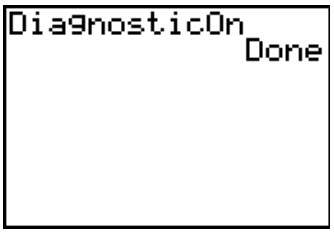


We're now here. We haven't gotten to the command yet, but we're close.

Press the down arrow ▼ until the command DiagnosticOn is highlighted. Press ENTER to select the command and transfer it to the home screen, then ENTER again to execute it.



Your screen should look like the one at right.



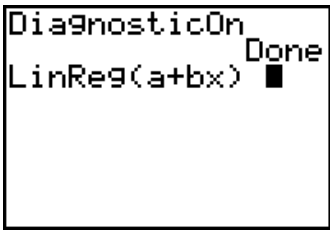
REGRESSION AND CORRELATION

We're now ready to examine the correlation between these two variables. However, the calculator will not give just the value of r; it's much easier computationally for it to do the whole thing at once and report all the statistics of interest.

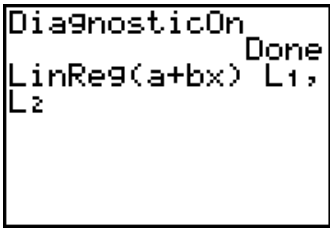
TI-83/84 Procedure

Press [STAT], arrow to CALC. (We've been here before for 1-var Stats). There are two linear regression choices: 4:LinReg(ax+b) and 8:LinReg(a+bx). The answers you get will be the same, but one must keep in mind the order in which the coefficients are used. Since statisticians usually prefer the constant term of the regression to come first (in case there are several predictor variables – multiple regression) we'll use choice 8.

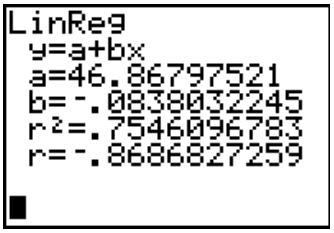
Either press the down arrow until the selection is highlighted then [ENTER] or simply press [8]. The command will be transferred to the home screen. In doing a simple regression with predictor variable in L1 and response in L2, simply pressing [ENTER] at this point is enough. However, if you want to store the equation (to see it on your graph for one reason), you need to specify the lists in which the data is stored; it's also just good practice to get into the habit, since you may want to use lists other than L1 and L2.



Press [2nd][1] (L1), then [,], then [2nd][2] (L2) followed by [ENTER] to execute the command. Before pressing [ENTER] the screen should look like the one at right.



Once the command is executed, you should see the display at right. Notice the first line of the results displays the type of regression in terms of *y* and *x*. Regression lines should never be reported in these terms, but the calculator does not know what variables you are working with. This is really an aid to remind you where the coefficients a and b go in the equation.

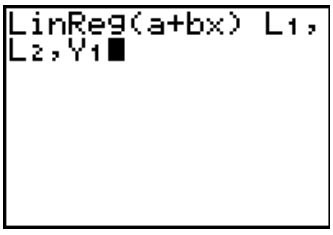


Here, we have the following regression equation: *EstimatedMileage* = 46.87 – 0.08**horsepower*. As always, how many decimal places to report can be subjective (and depend on the type of the data). For data like these, two decimal places should be sufficient. Ask your instructor for guidance. Remember, this line represents the average value of gas mileage for a given horsepower rating, based on the model from our data. The slope of -0.08 indicates that gas mileage decreases about 0.08 miles per gallon for each additional horsepower in the engine, on the average. In addition, we see the correlation coefficient, $r = -0.869$ which indicates a strong negative relationship. The coefficient of determination, $r^2 = 0.755$ (normally expressed as $r^2 = 75.5\%$) tells us that 75.5% of the observed variation in gas mileage (remember these values ranged from 10 to 31 mpg) is explained by the horsepower of the engine, using our model.

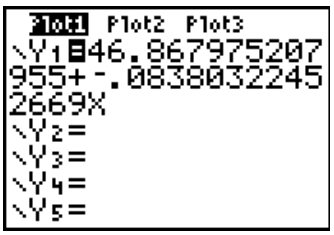
Storing the regression line

It was previously mentioned that the equation of the regression line can be stored for future reference. This is done by modifying the regression command as follows:

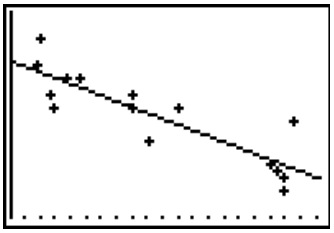
Press [STAT], arrow to CALC, then [8][2nd][1][,][2nd][2] (so far, this is what we did before). Now press [,][VARS] arrow to Y-Vars, press [ENTER] to select 1:Function, then [ENTER] to select Y_1. The regression command should look like that at right. Press [ENTER] to execute the command.



The regression output will look the same as before. The difference between the two commands can be seen by pressing [Y=]. The equation of the line has been stored for further use.

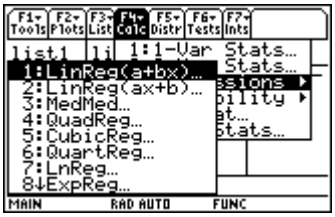


It would be nice to see how the line passes through the data; it should be roughly in the center of the data points. Press [GRAPH], since there is no need to resize the window. Sure enough, there's the line just as we expected. Notice that since no line will be perfect (unless $r = \pm 1$), some of the points are above the line, and some below. The vertical distances between the points and the regression line are called *residuals* and their plots are used to examine the line for adequacy of the model.



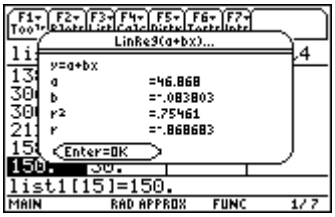
TI-89 Procedure

From the Statistics Editor, press [F4] (Calc). Arrow down to choice 3:Regressions and press the right arrow. Both choices 1 and 2 on this submenu are linear regressions. The answers you get will be the same, but one must keep in mind the order in which the coefficients are used. Since statisticians usually prefer the constant term of the regression to come first (in case there are several predictor variables – multiple regression) we'll use choice 1.
Press [ENTER] to select it.



You will be presented with an input screen like those we have seen before. It asks for the list containing the x (predictor) variable; the y (response) variable; and gives you the option of storing the equation of the line. With the right arrow here you can select none or a y-function. Since one usually wants to see the line plotted on the data graph, it is a good idea to select a function (usually y1(x)). Freq should be left at 1. My regression definition is at right. When finished with the definition, press [ENTER].

Once the command is executed, you should see the display at right. Notice the first line of the results displays the type of regression in terms of y and x. Regression lines should never be reported in these terms, but the calculator does not know what variables you are working with. This is really an aid to remind you where the coefficients a and b go in the equation.

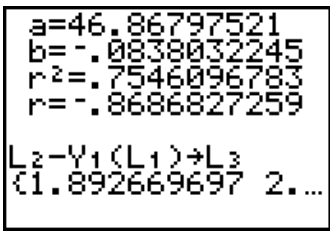


If you press [◆][F1] (Y=) you will see the regression equation as on a TI-83/84.
Pressing [◆][F3] (GRAPH) will graph the data with the line added.

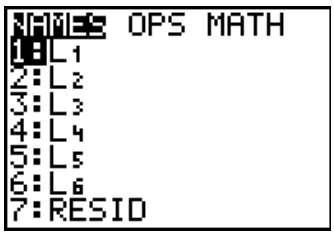
RESIDUALS PLOTS

Residuals are defined to be the vertical distance from the data point to the regression line, in other words, $e_i = y_i - (a + bx_i)$ for each data point (x_i, y_i) in the data set. The e_i are the residuals. There are two ways to obtain the residuals. The first makes use of the stored regression line. From the list of y-values we will subtract the value on the line by "plugging in" each corresponding x-value into the equation. The residuals will then be stored into a new list.

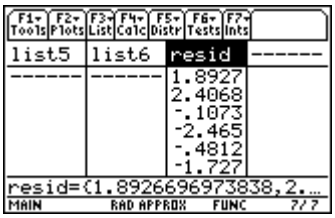
For our example we will enter the following command onto the home screen: 2nd 2 − VARS, arrow to Y-Vars, press ENTER to select Function, then ENTER to select Y1(2nd 1) STO▸ 2nd 3. This command says "take the *y*-list in L2 and from it subtract the value obtained by evaluating function Y1 at each *x* value in L1, then store the results into new list L3. Your command should look like the one at right. I have already carried out the command; the first few residuals are displayed. More can be seen by pressing ▸ or by using the STAT editor.



Alternatively, the calculator automatically finds residuals. They can be accessed with 2nd STAT (LIST). Notice there is a list called RESID. (You may have many more list names showing on this screen, depending on how the calculator has been used in the past). These are the residuals from the *last* regression performed.



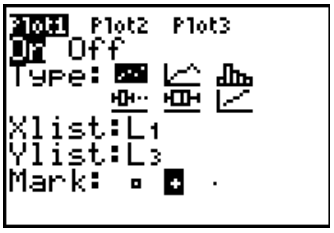
TI-89 calculators automatically add the list of residuals into the Statistics editor when the regression is calculated. When trying to find these in [VAR-LINK] to define plots, they will be in the STATVARS folder. You can use the right and left arrows to expand and hide folder contents. To locate the list more readily, in the STATVARS folder, press alpha 2 = R to move to the first variable that begins with the letter R (the correlation coefficient).



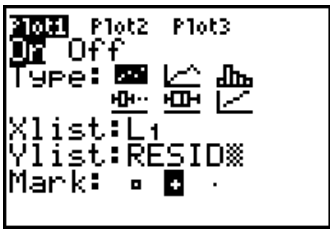
There are two main types of residuals plots which should be done to examine the adequacy of the model for any regression. The first plots the residuals against *x* (the predictor variable); the second is a normal probability plot of the residuals. In the first plot, we hope to see random scatter in an even band around the *x*-axis ($y = 0$ line). Any departures from this are cause for reexamination of the model. In particular, curves may appear which are "masked" by the original scaling of the data; subtraction of any linear trend will magnify any curve. Another common shape which indicates problems is a "fan" in which the plot either narrows from left to right or, alternatively, thickens. Either of the fan shapes means there is a problem with an underlying assumption, namely that the variation around the line is constant for all *x*-values. If this is the case, a transformation of either *y* or *x* is usually necessary. Unusual observations (outliers) may also be seen in these plots as very large positive or negative residuals. Plot definitions are similar for all calculators, with the exception of normal plots on the TI-89 (see the last chapter).

A residuals plot against *x* (the predictor variable)

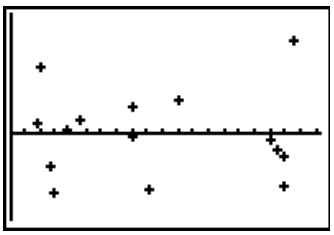
This is a scatter plot. From the plot definitions screen (press 2nd Y= ENTER to define Plot1) define the plot using the original Xlist of the regression (in our example L1) and the residuals list (from our example L3). Press ZOOM 9 to display the graph.



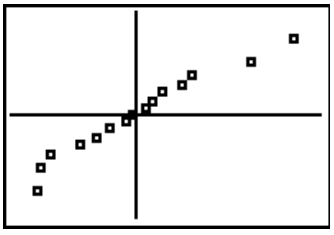
Alternatively, if using the automatic residuals, define the plot as above, but for the Ylist, press 2nd STAT and ENTER to select RESID, followed by ZOOM 9. The definition screen will look like the one at right.



In either case, the residual plot should look like the one at right. Looking at the plot, we see no overt curves, indicating a line appears to be an adequate model. This was a small data set; with these seeing non-constant variation can be difficult. There does not seem to be much of a problem except at the far right end of the graph, where there might be a high outlier, but it's hard to tell.



Normal Probability Plots of Residuals

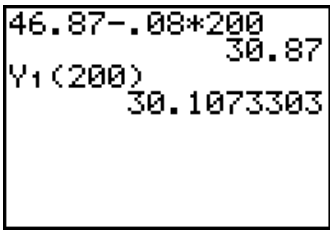


The second plot which should be done is a normal probability plot, since there is an underlying assumption the residuals have a normal distribution. This assumption will be used later in inference for regression. Normal probability plots were discussed in Chapter 4 of this manual. Remember that the data list is the list of residuals. We're looking for an (approximately) straight line. Here, the pattern is fairly linear, indicating no severe problem with this assumption.

Residuals Plots against Time

If the data were gathered through time (the data in our example were not) a time plot of the residuals should be done as discussed in Chapter 2. Ideally, this should look like random scatter. Any obvious patterns (lines, curves, fans, etc) indicate time is an important factor and the model which was fit is not adequate to fully describe the relationship. This generally means a multiple regression is needed to explain the response variable.

USING THE EQUATION TO PREDICT



Often we want to predict a value for the response variable based on the regression. This can be done in either of two ways: We can "plug" the value for the predictor (x) variable into the equation explicitly, or with the equation stored as a y-function, simply have the calculator evaluate the result. On the screen at right, we want to know the average gas mileage for a car with a 200-horsepower engine. Notice that since one tends to round the reported slope and intercept, the two answers might disagree. It's much better to have the function evaluate the desired amount using the largest number of significant figures, then round the final result. Based on our linear regression, the average highway gas mileage for cars with 200 horsepower is about 30.1 miles per gallon. We had two cars in our data set that had 200 horsepower: the Audi (32 mpg) and the Buick LaCrosse (30 mpg).

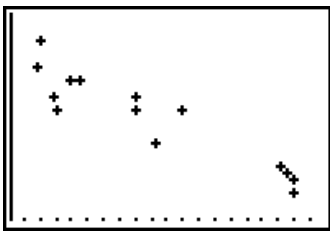
To locate the `Y1` function, on a TI-83/84 press [VARS], then arrow to `Y-Vars`, press [ENTER] to select `Function`, then [ENTER] again to select `Y1`. On a TI-89, use [VAR-LINK] to find the variable in the `Main` folder.

IDENTIFYING INFLUENTIAL OBSERVATIONS

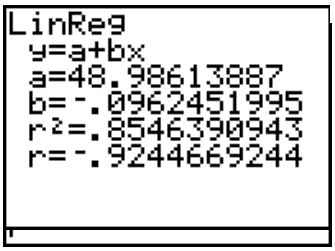
Remember, the large residual on the far right side of the plot looked unusual. Points far away from the center of the range of the predictor variable can be influential; that is, they may have a significant impact on the slope, especially if they do not follow the pattern of the rest of the data. Even if they do not impact the slope, they will cause r and r^2 to be larger than the rest of the data would warrant. To decide if points are influential, delete the suspects, and reanalyze the data.

The large residual corresponded to the Lexus IS 350. It had the highest horsepower rating of all our vehicles, and had a higher highway gas figure than any other vehicle with a similar horsepower rating. Let's delete that data point (be sure to delete both the x and y values from each list) and see if it is influential in some way.

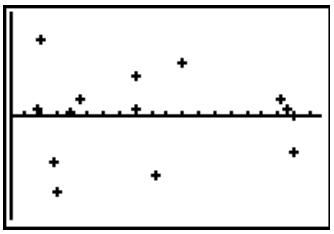
Redrawing the scatterplot (L1 as Xlist and L2 as Ylist) with ZOOM 9 gives the plot at right. This looks much more linear than the original data scatterplot. One now could suspicious of the cluster of big horsepower vehicles at the lower right, but they're real data!



The new linear regression output is at right. The new regression equation is *Mileage* = 48.97 – 0.10**horsepower*. The original equation was *Mileage* = 46.87 – 0.08**horsepower*. The slope changed by about 20%, which is a fair amount. Both *r* and r^2 increased as well. The Lexus may well be influential.



Here's a residuals plot against horsepower (*x*). This plot is indicative of another problem that might be encountered in a regression – there is less variability in the residuals as we go from left to right on the graph. We have a violation of the equal variance condition.



We are left with the following indications: the Lexus was influential on the regression slope. Removing it from the data set impacted the slope (and intercept and *r* and r^2). However, because of the residual plot above, the indication is that a line is *not* the proper model to describe the relationship between horsepower and gas mileage. What is correct? Perhaps some type of transformation of the data will be better.

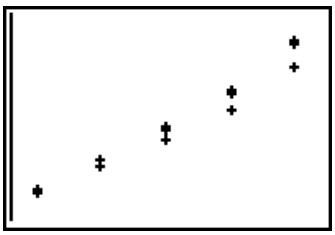
TRANSFORMING DATA

There are two reasons to transform data in a regression setting: to straighten a curved relationship and to transform variability so it is constant around the line. In a single variable case, transformations can be used to make skewed distributions look more symmetric; in the case of a single variable observed for several groups, a transformation can make the spread of the different groups look more equal.

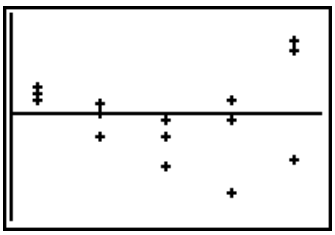
The table below shows stopping distances in feet for a car tested three times at each of five speeds. We hope to create a model that predicts stopping distance from the speed of the car. (Data are from problem 17 in Chapter 10 of the text.)

Speed (mph)	Stopping Distance (ft)
20	64, 62, 59
30	114, 118, 105
40	153, 171, 165
50	231, 203, 238
60	317, 321, 276

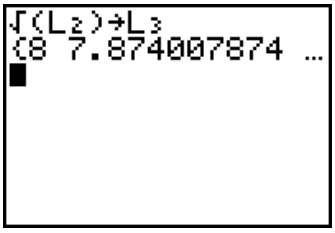
A plot of the data is at right. It looks fairly linear, but it is clear that the stopping distances become more variable with faster speed.



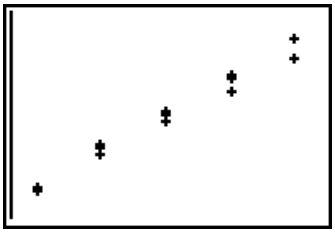
Regression gives the fitted model as *Stoppingfeet* = −65.933+5.98**Speed*. The residuals plot against speed (at right) clearly indicates the variability gets larger for faster speeds; it also indicates the true relationship is not linear but curved. Clearly, a transform is indicated – one which will decrease variation as well as straighten the plot.



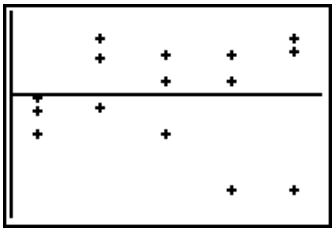
Since the residuals indicate a curve (possibly quadratic), using the square root of stopping distance makes sense. With stopping distance in L2, we need to find the square root of each distance. With one command we can do this, storing the result in a new list, say L3. Press 2nd x^2 ($\sqrt{}$) 2nd 2) STO▶ 2nd 3 followed by ENTER. The command and result are at right. We see the first couple values. How many are shown depends on the number of decimal places displayed. To see the entire list, go to the STAT editor.



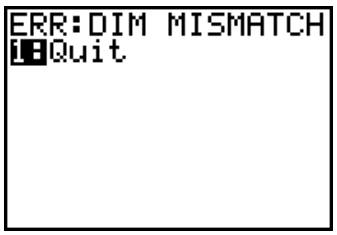
The new scatterplot is at right. The new regression equation is *sqrt*(*Stoppingfeet*) = 3.303 + 0.235**Speed* We have $r = 0.9922$, an extremely strong linear relationship. What about a residuals plot?



Here is the residuals plot. It's not perfect; the variation still increases with larger values of speed, but is much better than before. Sometimes there is no "perfect" transform.

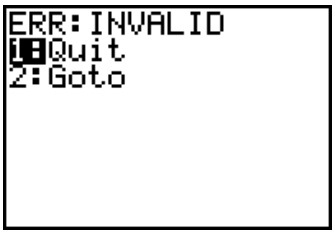


WHAT CAN GO WRONG?



What's Dim Mismatch?

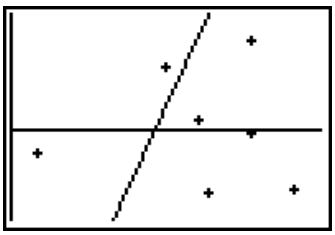
We've seen this one before. Press ENTER to quit. This error means the two lists referenced (either in a plot or a regression command) are not the same length. Go to the STAT editor and fix the problem.



What is Err: Invalid?

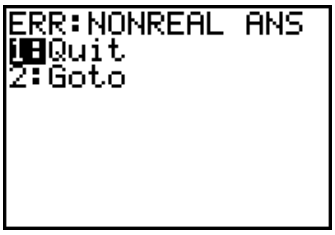
This error is caused by referencing the function for the line when it has not been stored. Recalculate the regression being sure to store the equation into a *y* function.

What's that weird line?



This error can come either in a data plot (an old line still resides in the Y= screen) or the stored regression line is showing in the residuals plot, as shown here. The regression line is not part of the residuals plot and shows only because the calculator tries to graph everything it knows about. Press Y= followed by CLEAR to erase the unwanted equations, then redraw the graph by pressing GRAPH.

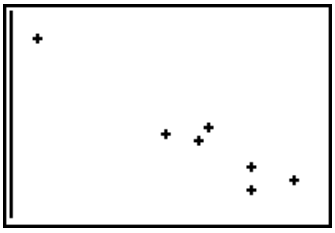
Audi A4	200 hp	32 mpg	Honda Accord	166	34
BMW 328	230	30	Hyundai Elantra	138	36
Buick LaCrosse	200	30	Lexus IS 350	306	28
Chevy Cobalt	148	32	Lincoln Navigator	300	18
Chevy Trailblazer	291	22	Mazda Tribute	212	25
Ford Expedition	300	20	Toyota Camry	158	34
GMC Yukon	295	21	VW Beetle	150	30
Honda Civic	140	40			



What does Nonreal Ans mean?

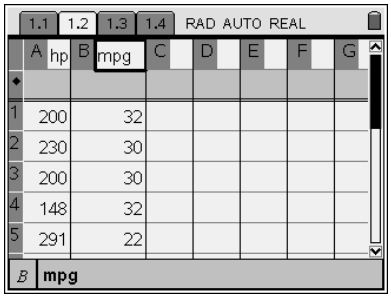
This error comes from trying to take the square root (or log) of a negative number. This can't be done in the real number system. These transforms do not work for negative values. Try something else.

This doesn't look like a residuals plot!

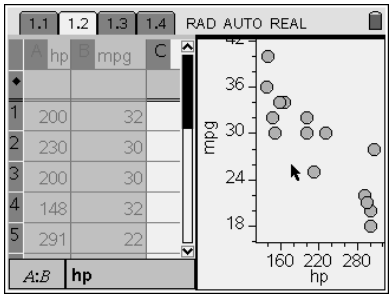


It doesn't. Residuals plots *must* be centered around $y = 0$. This error is usually caused by confusing which list contains the y's and which the x's in finding residuals by "hand." Go back and check which list is which, then recomputed the residuals, or use the list automatically stored by the calculator (under the LIST menu.

Commands for the TI-Nspire™ Handheld Calculator

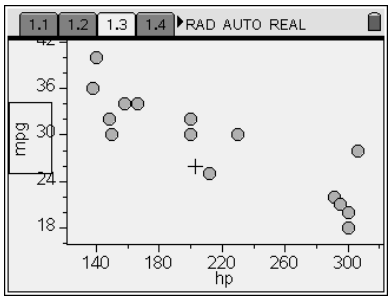


To create scatterplots and regression lines, first enter the data as lists. Name each list. To use the data in the following table (the same data shown earlier in the chapter), create two lists, named *hp* and *mpg*.

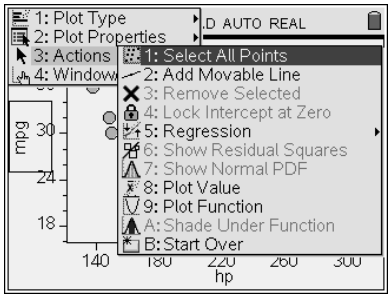


With the cursor in the first column, press ▲ until the column is highlighted. Then press ⇧ and then ▶ so both columns are highlighted. For a scatterplot on a split screen, press menu , select Data, and then Quick Graph.

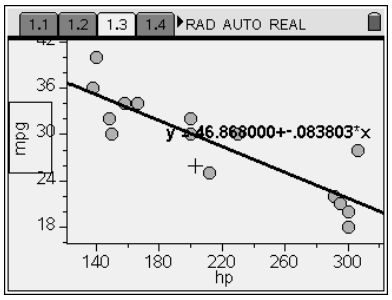
To create a scatterplot on a full page, press (home) and then select Data & Statistics. At first you will see a scatterplot, but not the one you expected. Use the arrows to move to the bottom of the display until "Click to add variable" appears. Press (ctrl) , highlight the variable, in this case *hp*, and press (enter). A dotplot appears. Now repeat the process to use *mpg* on the y-axis.



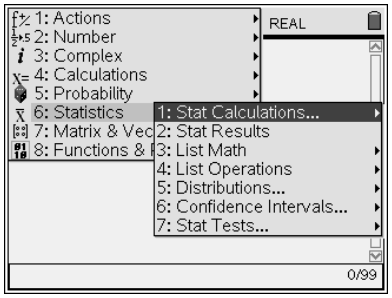
To display the least-squares regression line, press (menu) , and then select Analyze , Regression, and Show Linear (a+bx). The line and the equation are displayed.

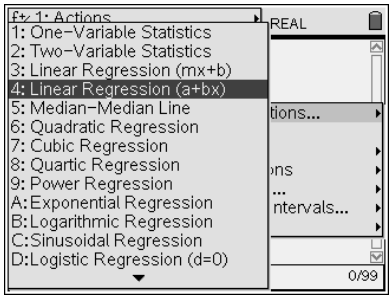


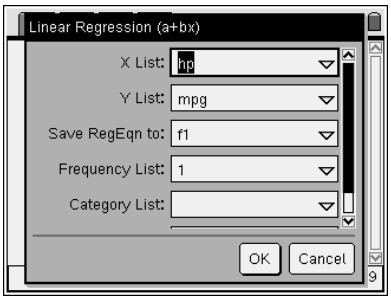
.

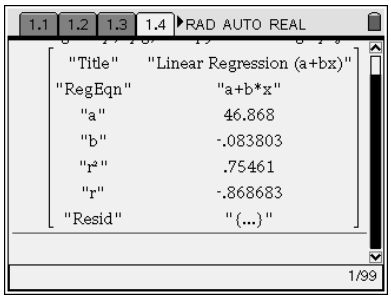


You may also obtain the equation for a least-squares regression line from a Calculator page. Press (menu) and select Statistics, then Stat calculations, and Linear Regression (a+bx). In the input box, select the list names, (tab) to OK and (enter).









To display a residual plot, open a Data and Statistics page, and select the variable stat.resid for the y-axis.