

> What Is a Linear Regression?

You can find the equation that best fits a set of data points!
Combined with the Texas Instrument Calculator Based Laboratory 2 ${ }^{\text {TM }}$ System (CBL 2 ${ }^{\text {TM }}$ ), the Voyage ${ }^{\text {TM }} 200$ PLT allows you to almost effortlessly fit a variety of different curves to data collected in real time. The resulting regression equation of the curve is then said to model the data. A regression equation model almost never models real data perfectly, but no matter what model you choose-linear, quadratic, cubic, quartic, exponential, power, or logarithmic, to name some of the regressions built into the Voyage 200 PLT-the mathematics for finding the best-fit regression equation is interesting in itself.

This exercise consists of two parts. In part one, you will create lists of data points, plot them on a graph, and then let the Voyage 200 PLT determine the regression equation of the line that best fits the data points. In part two, you will use the Voyage 200 PLT to understand exactly how the linear regression is determined.

## Instructions

## Part A—Finding a Linear Equation

Find a linear equation whose graph best fits the five points $(1,6),(2,10),(3,13),(4,12),(5,17)$.

1. Press 00 to turn on the Voyage 200 PLT. To reset to the default settings, press [2nd [MEM] F1, select 1:RAM, select 2:Default, and then press ENTER [ENTER. Press [CALC HOME] to access the Home screen. Delete all values of one-character variables by pressing F6, selecting 1:Clear a-z, and pressing ENTER.
On the Home screen, clear the entry line by pressing CLEAR once or twice. Clear the History area above the entry line by pressing F1 8 .
2. To begin, you need to make a list of the $x$-coordinates and store it under the variable name xs. At the entry line of the Home screen, press
2nd [i] 1
$\square$
$2 \square$ $\square$ 4 5 [2nd [ 1 ] STO. XS ENTER.

Make a list of the $y$-coordinates, and store it under the variable name ys. Press
2nd [i] 6
10 $\square$ 13 12 $\qquad$ YS ENTER.

## What Is a Linear Regression? (Continued)

3. To set up a statistical data plot, press $\sim[\gamma=]$ to display the $\mathrm{Y}=$ Editor. If there are any functions or stat plots there, press F5 and select 1:All Off to deselect them.

Press $\Theta$ to select the first empty plot above $\mathrm{y} 1=$ and press ENTER. Plot Type and Mark are set to the desired defaults.

Scroll down with $\odot$ to the box $x$ and type XS. Then scroll down to $y$ and type YS. You will see the display at the right. Press ENTER ENTER to save.
4. To split the screen, return to the Home screen by pressing [CALC HOME] and change the screen mode to split by pressing MODE F2 (1) 3 ENTER. The screen is now split with the Home screen on the left and a graph screen on the right.
Graph the data points by pressing $\bullet$ [GRAPH]. (The graph window is not set to display all the data points.) To fit the data points on the screen, press (F2 and select 9:ZoomData. You should see the display at the right.
5. Notice that the graph screen on the right has a black border around it. It is the active screen. Activate the Home screen on the left by pressing 2nd [ $\boxplus$ ]. (You can "toggle" back and forth between screens in split screen mode by using 2nd [ $\boxplus$ ].)


## What Is a Linear Regression? (Continued)

6. You now need to define a linear equation of the form $y=m x+k$. You could do this in the $\mathrm{Y}=$ Editor, but instead, do it directly from the Home screen as follows:

$$
\text { Press } \mathbf{M} \boxtimes \mathbf{X} \boxplus \mathbf{K} \text { STO Y1 } \square \mathbf{X} \square \text { ENTER. }
$$

Verify that you have the right equation by pressing Y1 $\square \mathbf{X} \square$ ENTER.

You should see as your answer $m x+k$. If, and only if, you see numbers instead of $m$ and $k$, you did not delete values for the variables; in which case, press F6 ENTER and repeat step 6.
7. At this point, you need to operate on the data lists to define a numerical measure of good fit, ss, used in the Method of Least Squares. The idea behind the method is this:

If a line fits the points well, the distances of the points from the line are going to be as small as possible. Because you don't want the positive differences of points above the line to be canceled out by the negative differences of points below the line, you square the differences, making the squares of the distances between the points and the line positive. You then add the squares of these distances and consider the line to fit the points best if the sum of the squares of the distances is a minimum.

Press F4 and select 1:Define. Then press
$\mathbf{S S} \square \mathbf{S U M} \square \square \mathbf{Y S} \square \mathbf{Y 1} \square \mathbf{X S} \square \square \mathbf{2} \square$ ENTER.

Verify that the formula is correct by pressing
SS ENTER. You should see the first part of the expression
$5 k^{2}+k(30 m-116)+55 m^{2}-396 m+738$.
(To see the rest of the expression, press $\Theta$ and then press (1).) If you do not see this expression, repeat step 7 .

## What Is a Linear Regression? (Continued)

8. Guess values for $m$ and $k$ for a line to fit the data points. For an initial guess, choose $m=2$ for the slope and $k=2$ for the $y$-intercept. Then graph the line.

Press ENTER CLEAR
2 STOD M ENTER 2 STOص K ENTER [GRAPH].
You should see the display on the right. It is not a particularly good fit.

Press [2nd [ $\boxplus$ ] to return to the Home screen. Now use the equation ss to measure the fit by pressing SS ENTER.

Because you assigned values to $m$ and $k$, you should see the number 74 .
9. With $m=2$ and $k=2$, the line $y=2 x+2$ had a sum of squares of 74 . Record those values in the first line of the table below.

From the graph of the line, you can tell that both the slope and $y$-intercept were a bit too low. Estimate new values of $m$ and $k$ for a line that better fits the data points by pressing
(new guess for $m$ ) STO• M ENTER
(new guess for $k$ ) STOD K ENTER
SS ENTER.
Record your values in the table below, and display the graph by pressing [GRAPH] to see the new line.

If your guess was an improvement, not only should the line look closer but the value of ss should be smaller.

Repeat this step several times until you are satisfied you have a line that appears to fit the data well. (Remember to toggle back to the Home screen by pressing [2nd [ $\ddagger$ ] after each new graph.)

| $m=$ | $k=$ | $y=m * x+k:$ | ss = |
| :---: | :---: | :---: | :---: |
| $m=$ | $k=$ | $y=m * x+k:$ | ss $=$ |
| $m=$ | $k$ | $y=m * x+k:$ | ss = |
| $m=$ | $k=$ | $y=m * x+k:$ | ss = |

## What Is a Linear Regression? (Continued)

10. The Voyage ${ }^{\text {TM }} 200$ PLT uses the Method of Least Squares to calculate the line of best fit. From the Home screen, press [2nd [MATH] and select 6:Statistics, 3:Regressions, and 1:LinReg. Then press XS YS ENTER.

To see the results, type SHOWSTAT and press ENTER.
Notice that the regression equation is written with coefficients $a$ and $b$ instead of $m$ and $k$. Record the values below. Press ENTER and store the calculated values to $m$ and $k$, respectively. Press SS ENTER to calculate the value of ss. Graph by pressing [GRAPH].
$a=m=\quad b=k=\quad y=m * x+k: \quad$ ss $=$

The value of ss is the least possible value for the given data, hence the name, Method of Least Squares.

## Part B—Calculating an Algebraic Equation for SS

The Method of Least Squares approach of fitting a line to the points can also be analytical, as opposed to visual (although the visual approach works surprisingly well).

Examine the values of ss from your table above. Notice that the smallest value of ss is associated with the regression line. Think of it this way-if the data points all lay exactly on a line, the value of ss would have to be zero!

Now you are going to calculate an algebraic expression for ss in terms of unknown values of $m$ and $k$.

1. To return to the Home screen and quit split screen mode, press [2nd [QUIT] from the Graph screen.

Delete current values of $m$ and $k$ by pressing
F4 and selecting 4: DelVar. Then press M, K, ENTER.
Display the value of ss with undefined values of $m$ and $k$ by pressing SS ENTER. Once again, you should see the expression
$5 k^{2}+k(30 m-116)+55 m^{2}-396 m+738$.
Notice that ss is quadratic in both $m$ and $k$.
Recall that in the quadratic function $y=a x^{2}+b x+c, a$ is the coefficient of the quadratic term and $b$ is the coefficient of the linear term.

## What Is a Linear Regression? (Continued)

1. (continued)

Recall from algebra that the graph of $y=a x^{2}+b x+c$ is a parabola opening upward when $a>0$. Recall that the minimum value $y$ of the function occurs at the vertex of the parabola.

Thus, a quadratic function in $x$ has a minimum value when $x=\frac{-b}{2 a}$.
2. To minimize the value of $s s$ with respect to $m$, start by storing $x$ into $m$.

Press X STO^ M ENTER SS ENTER to redisplay the expression.

Identify and record the expressions corresponding to $a$ and $b$.
$a=$ $\qquad$ $b=$ $\qquad$ .
Then,
$-b / 2 a=$ $\qquad$ is the value of $m$ to minimize ss. Type these recorded expressions, and store them into $a$ and $b$, respectively.

Delete the current value of $m$ by pressing F4, selecting 4:DelVar, and pressing M ENTER.

Enter and store the first equation, which must hold true to minimize ss, by pressing

3. To minimize the value of ss with respect to $k$, start by storing $x$ into $k$, and then redisplay the expression.

Press X STO K ENTER SS ENTER.
Identify and record the expressions corresponding to $a$ and $b$ :
$a=$ $\qquad$ $b=$ $\qquad$
Then,
$-b / 2 a=$ $\qquad$ is the value of $k$ to minimize
ss. Type these recorded expressions, and store them into $a$ and $b$, respectively.

## What Is a Linear Regression? (Continued)

3. (continued)

Delete the current value of $k$ by pressing [F4, selecting 4:DelVar, and pressing K ENTER.

Enter and store the second equation, which must hold true to minimize ss, by pressing

4. You now need to solve the system of two equations in $m$ and $k$. Because eq1 expresses $m$ in terms of $k$, and eq2 expresses $k$ in terms of $m$, solving the system is easy. Using a substitution method, it takes two steps:
(a) Solve the first equation for $m$, given that the second equation expresses $k$ in terms of $m$.

Press
SOLVE [EQ1 $\square$ M [ Znd [I] EQ2 $\rightarrow$ ENTER].
Record your answer below.
$m=$
(b) Store the value for $m$.
(value of $m$ ) STO• M ENTER
Solve the second equation for $k$, given that the first equation expresses $m$ in terms of $k$.

Press

Record your answer below.
$k=$
Do these values agree with the regression coefficients found in step 6 of Part A?

## What Is a Linear Regression? (Continued)

## Extra Practice

Six students in a calculus class were given Voyage ${ }^{\text {TM }} 200$ Personal Learning Tools to help them improve their grades. The table below compares the number of hours each student practiced with the Voyage 200 PLT and his or her individual exam grade.

Assume that there is a linear correlation between the hours practicing with the Voyage 200 PLT and the grade. Use Part $A$ of the previous activity as a model to find the values of $m$ and $k$ that produce the least sum of squares.

Check your solution by using the LinReg( command on the hours and the scores to calculate the linear regression directly.

| hours worked with <br> Voyage ${ }^{\text {TM }} \mathbf{2 0 0}$ PLT | score on calculus <br> exam |
| :---: | :---: |
| 4 | 62 |
| 9 | 99 |
| 2 | 48 |
| 5 | 64 |
| 7 | 85 |
| 8 | 92 |

## What Is a Linear Regression? (Continued)

## Extensions

## The Exponential Regression

An exponential equation has the form $y=a * b^{x}$. If you take the natural logarithm of both sides of this equation, the result is $\ln y=x * \ln b+\ln a$. This equation is linear; that is, $\ln (y)$ is a linear function of $x$. Therefore, you can perform a linear regression on the values of $x$ and the values of $\ln (y)$.

Using the data from the table below, define lists for months and garbage. Define a third list as the natural logarithms of garbage. (Press LN GARBAGE STO• LNGARB ENTER.)

Perform a linear regression on months and Ingarb. (Press LINREG [space bar]MONTHS $\square$ LNGARB ENTER.)

Exponentiate the regression coefficients.
(Press 2nd [ $\mathrm{e}^{x}$ ] REGCOEF) ENTER). Compare this result with that obtained by performing an exponential regression directly on lists months and garbage. (Press EXPREG [space bar] MONTHS $\square$ GARBAGE ENTER SHOWSTAT ENTER.)

You have calculated an exponential regression.

| pounds of garbage in Sam Slobvik's dorm room <br> after months away from home |  |
| :---: | :---: |
| months | garbage |
| 1 | 10 |
| 2 | 27 |
| 3 | 50 |
| 4 | 72 |

## A Program for a Classroom Activity

See Appendix A for a listing of the program BestFit(). This program allows you to explore linear regression problems by interactively fitting a line to data points.

