Linear Regression I

Ma. Louise De Las Penas, Phd
Ateneo de Manila University
Philippines

LEVEL
High School and University

OBJECTIVES
To use the calculator to
a) find a linear model that describes a real world problem given a set of data
b) analyze the given problem and predict certain outcomes given the linear model

Corresponding eActivity
linreg.g1e

OVERVIEW
A mathematical model describes a real-world problem in mathematical language. The objective of curve fitting, which is one aspect of mathematical modeling, is to find a relation that fits as well as possible to a given set of data. When the relationship obtained is linear, we call the process as linear curve fitting or linear regression.

EXPLORATORY ACTIVITIES
[Note] We shall use small letter *x* instead of capital X as shown on the calculator throughout the paper.

Activity 1: Consider the data in Table 1 representing the value of the *Vanguard Index Trust-500 Portfolio* for 1987 to 1997.

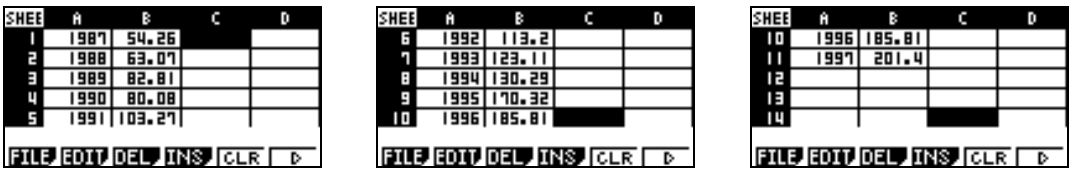
(a) Let *x* represent the year and *y* the value of the *Vanguard Index Trust-500 Portfolio*. Plot a scatter diagram for the given data.
(b) Compute the slope of the line joining the two points representing the Vanguard Index from 1987 to 1991. Do the same for the points representing the data from 1991 to 1995.
(c) Which slope in (b) is higher in magnitude? What does this imply?
(d) Find the line that best fits the given data. What is this line? Interpret the slope of this line. How best does this line fit the given data?
(e) If you were managing this trust, which of the three slopes would you use to convince

someone to invest?
(f) What is the trend in the data? As time passes what is happening to the *Vanguard Index Trust 500 Portfolio*?
(g) Assuming that the trend continues, what would be the value of the portfolio in 2006?

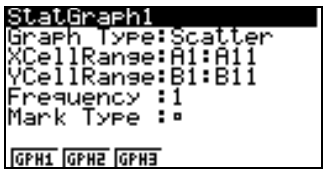
YEAR	VALUE (DOLLARS)
1987	54.26
1988	63.07
1989	82.81
1990	80.08
1991	103.27
1992	113.20
1993	123.11
1994	130.29
1995	170.32
1996	185.81
1997	201.40

Solution:

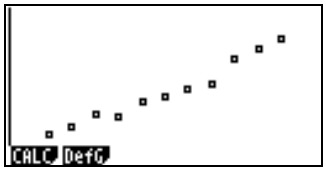
We open the Spreadsheet Editor and enter the given data. We enter the year in the first column and the value of the portfolio in the second column as follows:



(a) We press GRPH and SET the cell ranges for purposes of the scatter diagram:

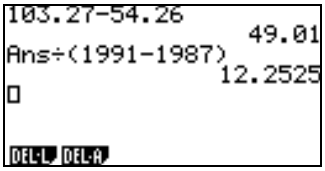


The scatter diagram of data is given as follows:

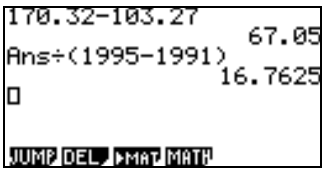


(b) We compute the slopes of the lines in the Run Editor. The slope of the line connecting

the points (1987,54.26) and (1991,103.27) is calculated as:

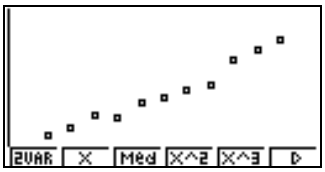


The slope of the line is 12.2525. On the other hand, the line connecting the points (1991, 103.27) and (1995,170.32) has slope 16.7625 calculated as follows:

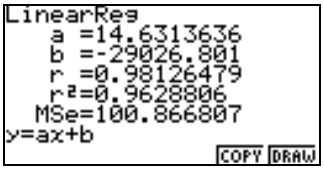


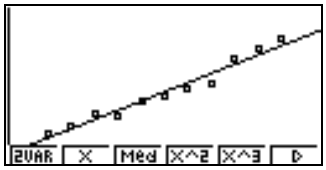
(c) The slope of the line connecting the points (1991, 103.27) and (1995,170.32) is higher in magnitude than the slope of the line connecting (1987,54.26) and (1991,103.27). This implies that between the years 1991 to 1995, the value of the portfolio increases by about 16.7625 dollars on the average. This is a higher value by about 4.51. (Between the years 1987 to 1991, the value of the portfolio increases by about 12.2525 dollars only on the average)

(d) To get the line of best fit, we enter the linear regression command(X) while the scatter plot is one the screen:



The line of best fit is given to be $y = 14.6313636x\text{-}29026.801$. The line is drawn on the scatter plot as follows:





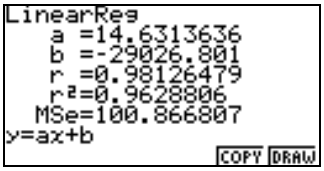
The slope of the line of best fit can be interpreted to mean that *x* increases by one year, the value of the portfolio increases by about 14.6313636 dollars.

How good is the line of best fit? The answer can be obtained by looking at the **correlation**

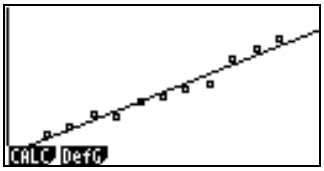
coefficient. The coefficient of correlation *r* is the measure of the strength of the linear relation that exists between two variables. The closer $|r|$ is to 1 the more perfect the linear relationship is. In the screen dump above, we see that *r* is 0.98126479, which is a good indication of a linear relationship between the variables.

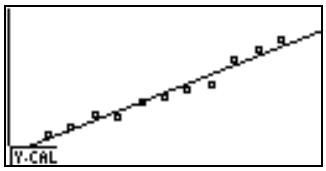
(e) It would be more reasonable to use the slope 14.6313636 of the line of best fit, which gives the increase in the value of the portfolio as *x* increases by a year and is an indication of the trend of the data.
(f) As mentioned in (e), the trend in the data is that the value of the portfolio increases as *x* increases.
(g) Assuming the trend continues, we can use the line of best fit to determine the value of the portfolio in 2006.

First, we copy the calculations we have obtained regarding the line of best fit to the graph menu within the Spreadsheet Editor.

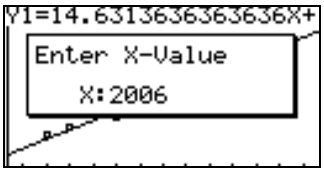


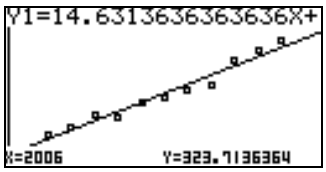
Next we draw the graph (left screen dump below). Then, pressing SHIFT G-Solv and Y-CAL will allow us to approximate the value of the portfolio given any year after 1997 using the line of best fit (right screen dump).





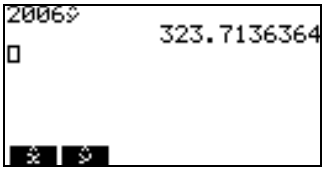
We can specify the x-value, for instance, *x*=2006, and the calculator gives the corresponding *y* value on the graph.





In year 2006, we approximate the value of the portfolio to be approximately $323.71.

The *y* value corresponding to *x* = 2006 using the line of best fit can also obtained from the Run Editor, using the OPTN STAT command as follows:



In the Run Editor, an *x* value corresponding to a given *y* value may also be obtained. For instance, a related question may arise as follows:

(h) Assuming the trend continues, when will the value of the portfolio reach approximately $500?

In the Run Editor, we obtain the following calculation:

Using the regression line, we approximate that the value of the portfolio would reach about $500 in the year 2018.

Remarks: It has to be noted that calculations in the Run Editor pertaining to data saved in the Spreadsheet Editor, (such as the calculations seen in the last two screen dumps above), cannot be performed while in eactivity mode. This can be performed only when using the Spreadsheet Editor from the Main Menu.

EXERCISE:

The following data gives the number of people reported as dying from AIDS in the United States from 1982 to 1992.

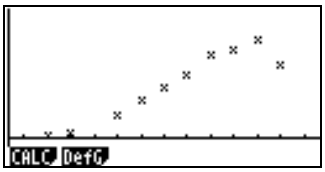
YEAR	1982	1983	1985	1986	1987	1988	1989	1990	1991	1992
Deaths	843	1651	6681	11535	15451	19656	26151	28053	30579	22660

(a) Based on these data, compare the interpolated number of deaths in 1990 with the reported number, 28053.
(b) Use the linear regression model to estimate the number of deaths from aids in the year 2006.
(c) If the trend continues, is it possible for the deaths from aids to reach about 200,000? At what year?

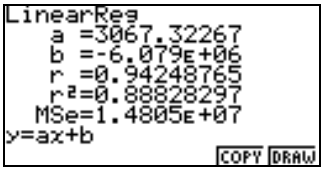
Solution:

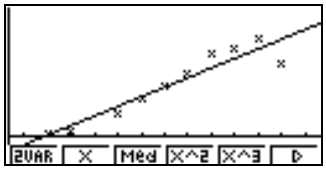
Linear Regression I

(a) In the Spreadsheet Editor, we enter the data and plot the scatter diagram.

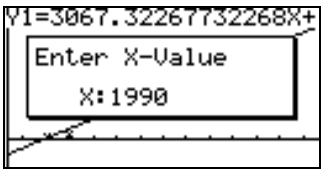


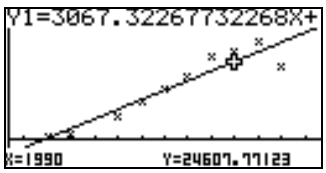
Then the line of best fit is obtained as follows, with the line drawn on the scatter plot.



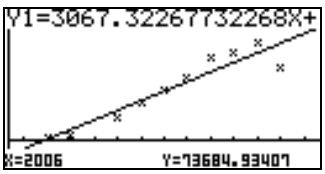


The regression line is given to be $y = 3067.32267x - 6079000$.Using this regression line, the interpolated number of deaths in 1990 is calculated to be 24,607.77. The reported number of deaths is 28053. A difference of approximately 3445. Note that the line we obtain has correlation coefficient, $r = 0.94$, though it suggests a high positive correlation, it is not a perfect one, so we expect some difference between the interpolated and true values.

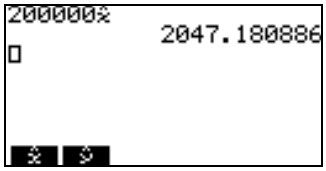




(b) The estimated deaths in 2006 is about 73,685.



(c)In the Run Editor, we approximate that in 2047, the deaths will reach about 200,000.



REFERENCE

[1] Sullivan, Michael. *PreCalculus*, 5th Edition. Prentice Hall Inc, USA, 1999.

