CASIO

Correlation Coefficient II

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

LEVEL
High School and University

OBJECTIVES
The aim of this note is to study and interpret the correlation coefficient corresponding to different types of regression with the aid of the graphics calculator.

Corresponding eActivity
correlat.g1e

OVERVIEW
Regression analysis is one of the branches of statistics that is used to compare quantities or variables, to discover relationships that exist between them and to formulate these relationships in useful ways. The first step in regression analysis is to graph the data using a scatter diagram and then to draw a curve that best fits the pattern exhibited by the data points. The best fitting curve for the sample points is called an **estimated regression curve**.

The **correlation coefficient** *r*, is the measure of how fitted the curve is to the data points. The closer $|r|$ is to 1, the better the correlation and the more the function best fits the data.

In this paper, we will illustrate how the graphics calculator is a helpful tool in regression analysis by providing the correlation coefficients for most types of regression curves.

EXPLORATORY ACTIVITIES
Example. Consider the following data regarding the number of farms in the United States over the years 1910 to 1999:

YEAR	1910	1920	1930	1940	1950	1959	1969	1978	1987	1999
Number of Farms (in millions)	6.4	6.5	6.3	6.1	5.4	3.7	2.7	2.3	2.1	1.9

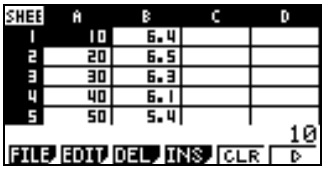
Correlation Coefficient II

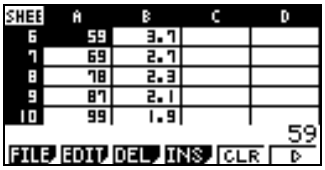
a. Draw a scatter diagram of the data.
b. Determine the function that best fits the given data.
c. Use the answer in b. to estimate the number of farms in 1900 and 1975.

Solution:

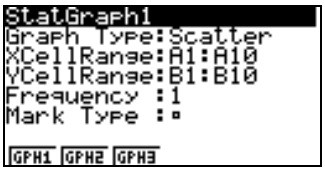
We access the Spreadsheet Editor.

Let the coordinate *x* of each data point be the number of years after 1900 and the coordinate *y* represent the number of farms. The year is entered in column *A*(1910 is entered as 10, 1920 as 20 and so on) and the number of farms in millions is entered in column *B* as follows:

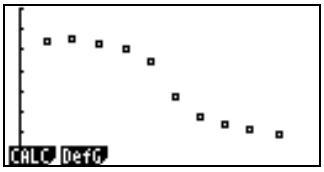


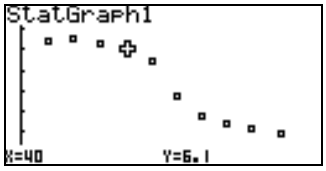


To display the scatter diagram for the given data, we access the GRPH menu. We assign the first graph, GPH1 to the given data. We specify the columns representing the *x* and *y* values and select " scatter diagram"[Scatter] as the graph type.



The scatter diagram is given below. Note that the scatter plot can be traced(right screen below) which is a good way of checking the data entries.

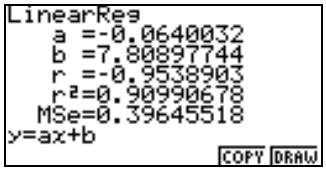




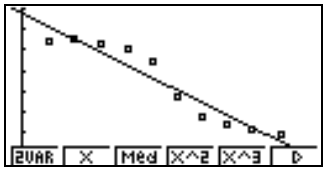
b. Once the scatter diagram has been drawn, the next step would be to explore the relationship between *x* and *y* and look for functions that fit the data approximately. We enter [CALC] and choose the type of function.

For instance, suppose we select the linear model. The regression coefficients are calculated as follows:

Correlation Coefficient II

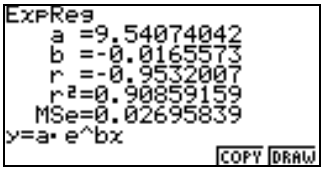


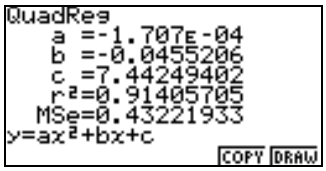
The linear model obtained is y =-0.0640032x + 7.80897744. Note that the correlation coefficient r is approximately -0.9538903. The correlation coefficient gives a measure of how well the data can be modeled by this linear function. Initially, we observe that the value $|r| \approx 0.95$ denotes a fairly good regression; but it is possible that a better fit is suggested by another functional model. In fact, if we generate the line on the scatter plot, it can be observed that the line does not seem to fit the data particularly well. There are some points that are not on the line, in fact, the scatter plot does not look so much like a line.

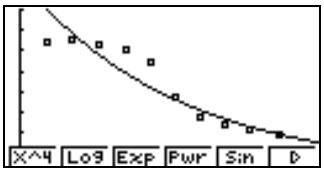


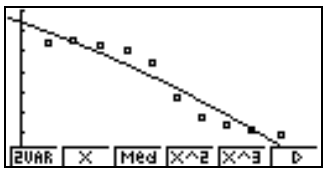
Thus, a best fitting curve of non-linear type will be sought. We experiment on other types of functions – quadratic, exponential, cubic, quartic, and look for a good fit to the data.

We have the following screen dumps for the exponential and quadratic regression:







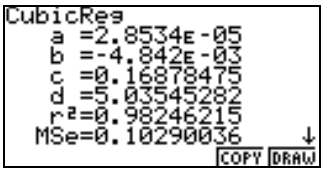


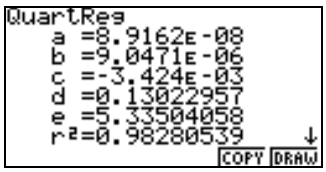
As in the linear model, we observe that there are some points that do not lie on the exponential and quadratic curves. The correlation coefficients corresponding to the exponential and quadratic models are approximated respectively as -0.9532007 and -0.9560633(about -0.95, the same approximation as that obtained in the linear model.)

Further exploration yields the following results: the cubic curve gives $r \approx$ - 0.9911923 and the quartic curve gives $r \approx$ - 0.9955864101. Both provide very strong correlation values: $|r| \approx 0.99$, which is very close to 1. However, the value of $|r|$ for the quartic curve is closer to 1 than that of the cubic curve. Thus, the best fit is apparently the quartic curve!

Correlation Coefficient II

The quartic function of best fit is given by the function
$0.000000089162x^4 + 0.000009047x^3 - 0.003424x^2 + 0.13022957x + 5.33504058$

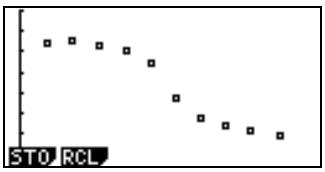




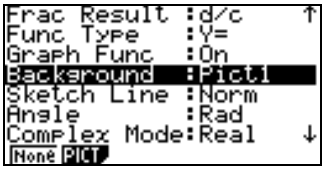
The negative *r* values obtained are indicative of the inverse nature of the relationship of the *x* and *y* values. As *x* increases, the *y* value decreases. This relationship can be visualized in the scatter plots. Note that the value of *r* may not be given directly, as in the case of the quadratic, cubic and quartic models. In which case, the square root of r^2 will have to be calculated, and the appropriate algebraic sign affixed depending on the relationship obtained, whether it is a direct or an inverse relationship. However, note that in verifying the function of best fit, it is not necessary to solve for *r*. One also can verify directly from the value of r^2, how well the function fits the data. The closer r^2 is to 1, the better the fit.

A good way to visualize the curve fitting process is to allow the calculator to draw more than one curve on the scatter plot. This is helpful particularly when comparing the graphs of the different functional models obtained relative to the data points.

A picture is obtained of the scatter plot and stored in the picture memory of one's choice, say Pict 1,by entering [OPTN]:

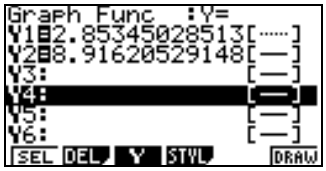


The picture is now used as background[SHIFT SET UP] as follows:

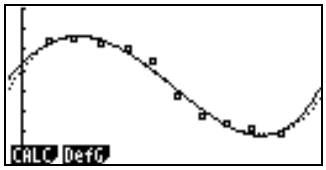


The functional models obtained after performing regression can be copied to the Graph function list and drawn with the picture as background.

For example, we copy the cubic and quartic functions obtained after regression to the Graph Editor. We assign dashes "----" to the cubic function and "__" to the quartic function. We access [DefG] to draw both graphs with the scatter plot as background:

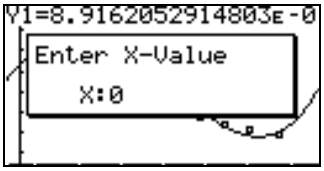


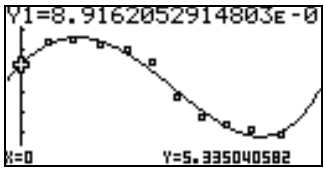
It can be visualized that both cubic and quartic curves fit the data points closely and confirm our earlier findings that both curves provide strong correlation values.

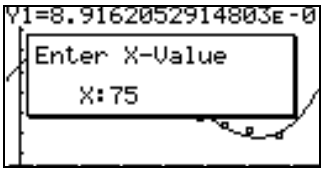


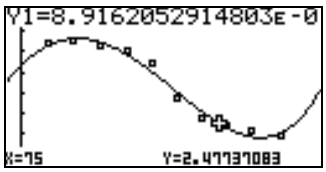
It is difficult to ascertain, at particular instances, which model to use over another by relying solely on the graphs or on the general shapes of the scatter plots. In deciding for example, which of the two models above to select, we refer to the value of the correlation coefficient to help us determine the function of better fit.

c. We use the quartic function obtained in b. to estimate the number of farms in 1900 and 1975. After entering [DefG], we press[DRAW], then [SHIFT G-Solv Y-Cal]. We specify the *x* value to obtain the corresponding value of *y*.





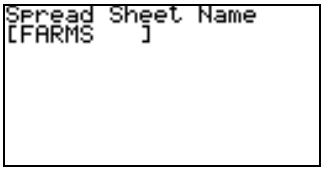


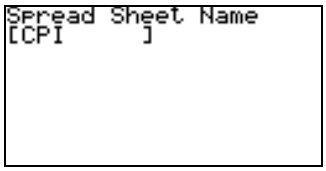


In Year 1900, there were about 5.335 million farms, and in 1975, there were about 2.477 million farms.

Remarks: After entering the data and other relevant settings during the activity, we save all information in the spreadsheet menu. We call the file name "FARMS'' and may be referred to at a later time. Similarly, for the exercise that follows, we call the file name "CPI".

Correlation Coefficient II





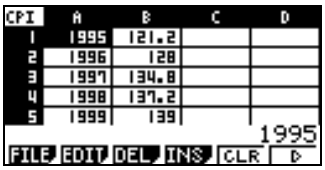
Exercise: Consider the following data indicating the levels of the Consumer Price Index(CPI) in December of the given year for food items:

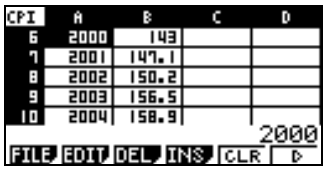
YEAR	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
CPI	121.2	128.0	134.8	137.2	139.0	143.0	147.1	150.2	156.5	158.9

a. Draw a scatter diagram of the data.
b. Determine whether an exponential model, logarithmic model, power model and linear model will best "describe" the relation between the year and the CPI.
c. Use the model obtained in b. to predict the CPI for food items in 2006.

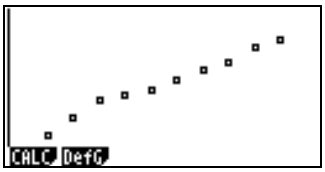
Solution:

a. We access the Spreadsheet Editor, and enter the years in the first column and the levels of CPI in the second column. We apply the sequence formula to generate the years by pressing [EDIT] then [SEQ].

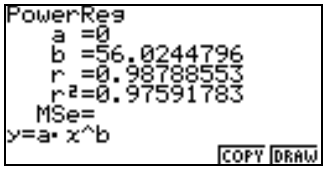


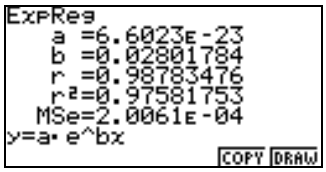


The scatter plot is given as follows:

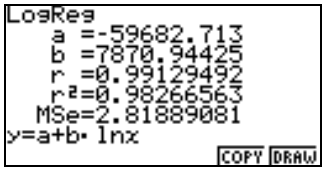


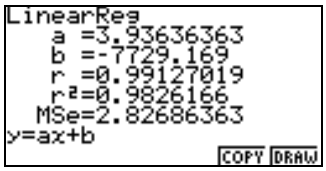
b. The regression coefficients for the various types of regression are given below:





Correlation Coefficient II





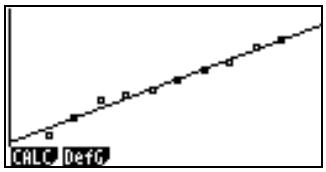
We summarize the results we obtained as follows:

CURVE TYPE	COEFFICIENT OF CORRELATION
Power	0.98788553
Exponential	0.98783476
Logarithmic	0.99129492
Linear	0.99127019

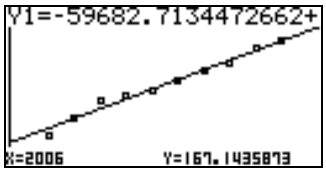
All four of the curve types provide a good correlation. The fact that *r* is positive indicates that there is a direct relationship between the *x* and the *y* values. As *x* increases then *y* increases. The value $|r|$ for all curves are close to 1, however, the best fit is apparently the logarithmic curve. The logarithimic model is given to be

$$-59682.713 + 7870.94425 \ln x.$$

The graph of the logarithmic curve superimposed on the scatter plot is shown below:



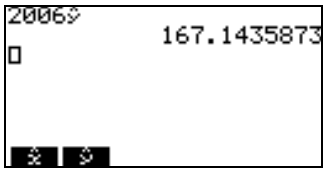
c. We use the logarithmic model as a predictor to determine the level of CPI in 2006. We obtain approximately 167.1435873.



Remarks:

The regression calculations and graphs obtained in the Spreadsheet Editor can also be obtained in the Stat Editor using the same commands. Moreover, when working outside the eactivity worksheet, it is possible to use the functional model to make calculations either through the Graph, Table and Run editors. To do this, the function must be first copied to the Graph Editor.

For instance, referring to Exercise c. given earlier, we calculate the value of *y* given *x* = 2006 in the Run editor as follows:

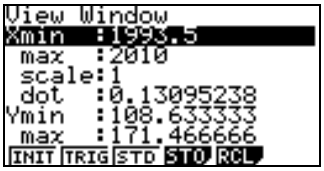


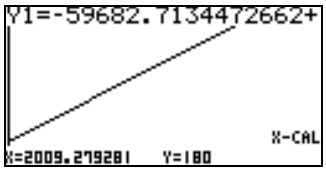
Working in either the Graph or Run Editor is helpful since not only can *y* values be determined, *x* values can be calculated as well.

Let us assume for instance that in the Exercise given above, we want to determine the approximate year when the CPI is 180.

Using the Run Editor, the answer given is year 2009:

In the Graph Editor, we obtain a similar answer using an appropriate view window:





REFERENCE

[1] Bittinger, et al. *Algebra and Trigonometry, Graphs and Models*, 2nd Edition, Addison Wesley, 2001.

