Linear Regression II

Jozef Hvorecky
Vysoká škola manažmentu / City University
Bratislava, Slovakia

LEVEL
High school or university students with basic knowledge in Mathematics.

OBJECTIVES
To use the calculator's built-in spreadsheet tool to make an elementary statistical evaluation – linear regression – for real-life problems.

Corresponding eActivity
S02GOLD.g1e (for Activity 1), S02FIELD.g1e (for Activity 2), S02HELTH.g1e (for Activity 3)

OVERVIEW
A collection of statistical data can be processed and used for predictions. The linear regression is one of the most popular methods.

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

Here we describe three activities. For their mathematical background refer for example to [LM], page 343-380.

Activity 1 (S02GOLD.g1e):

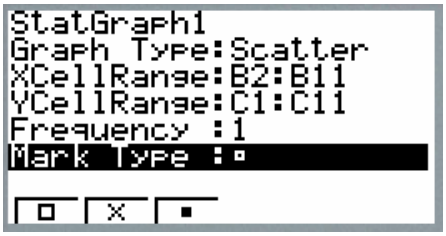
There is a gold-bearing river near a campus. The nearby gold mine has been exhausted many years ago, but diggers still occasionally pan the river sand for gold nuggets. Students created a Gold Digger Club as a sort of entertainment that could also to help them in collecting funds for their Christmas party. They go out on weekends, do panning and keep records on their achievements (see the spreadsheet table).

Linear Regression II

SHEE	A	B	C	D
1	Name	Days	Gold	
2	Ann	14	28	
3	Bob	35	66	
4	David	22	38	
5	Fiona	29	70	
6	Frank	6	22	
7	Kathy	15	27	
8	Mick	17	28	
9	Sarah	20	47	
10	Tim	12	14	
11	Xena	29	68	

As we see from the data, not all club members are equally active. Fiona – the club chair – believes that there is a relationship between the number of visits to the river and the amount of gold collected by the person. She wants to demonstrate it to the others. She is taking a Statistics course and wants to benefit from her knowledge of the evaluation method called Linear Progression.

(a) (Refer to Gold nuggets) Fiona enters data into a spreadsheet of her calculator. She first shows to the others that data can be displayed in a form of Scattered Points. To do the same, press [F6] and then [F1] (GRAPH). From its submenu select [F6] (SET).



The days spent on the river form the independent variable *x* (the cells B2 to B11), the amount of gold nuggets form the dependent variable *y* (the cells C2 to C11). Return to the graphic submenu and select [F1] – the graph is displayed.

The graph consists of isolated points – one for every digger. Even if they do not create an uncomplicated formation, there is a trend: For bigger x, the values of y tend to be bigger.

This trend can be expressed by a regression line. It has the general form

$$y = ax + b$$

where y represents the (approximate) amount of golden nuggets found in *x* days. The parameter *a* and *b* are calculated from the table values using the formulas:

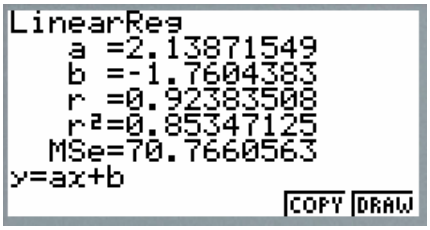
$$a = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - a\frac{\sum_{i=1}^{n} x_i}{n}$$

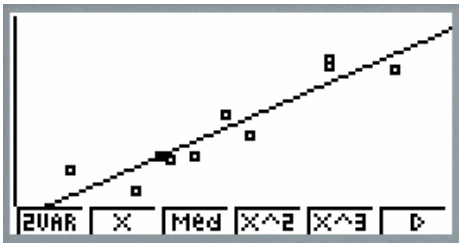
In general, we have to consider all *n* pairs of elements from two sets x_1, x_2, x_3, …, x_n (the number of visits to gold-bearing river per person) and y_1, y_2, y_3, …, y_n (the number of golden nuggets found by each individual). By combining the numbers in the prescribed method we get the result.

Fiona is not scared of these frighteningly-looking formulas as she knows that the calculator can evaluate them for her. To do it, press F1 (CALC). The new menu appears at the bottom of the display:

After pressing F2 (x), the calculation is performed and its result displayed.



The function that gives us the approximate values of gold nuggets after x days is therefore defined as y = 2.13871549x – 1.7604383. To view its graph, press F6 (DRAW).



EXERCISE A

Exercise 1.
How many golden nuggets we can expect to find after 30 days of panning?

SOLUTION to EXERCISE A

Exercise 1.
As we are making an estimate, we do not need to do very exact calculations. Two decimal places are sufficient:

$$(2.14 \times 30) - 1.76 = 62.4$$

We can expect finding about 62 nuggets.

Activity 2 (S02FIELD.g1e):
During the previous years, a farmer was using a fertilizer in his fields. He kept records on the amount of the used fertilizer (in tons) and yield (in tons of crops). The table shows his records.

SHEE	A	B	C	D
1	Year	Fert	Crop	
2	2000	2	71	
3	2001	1	36	
4	2002	3	79	
5	2003	4	104	
6	2004	2.5	73	

(a) (Refer to Scattered Graph) Using linear regression, show the dependence between the amounts of the fertilizer and the yields. Draw the scattered graph.

(b) (Refer to Calculation). Make the corresponding calculation. Record the parameters *a* and *b*.

(c) (Refer to Regression Line) Draw the regression line.

EXERCISE B

Exercise 1.
The farmer has made an option agreement with a bulk buyer on a planned purchase of 80 tons of crops from his next yield. How much tons of the fertilizer would you recommend the farmer to buy?

SOLUTION to EXERCISE B

Exercise 1.
The parameter *a* has the value 21.2, *b* has the value 19.6. We have to solve the equation

$$80 = 21.2x + 19.6$$

Its result is 2.84 (approximately 3 tons).

Activity 3 (S02HELTH.g1e):

On a sports event, a medical team measured the time achieved by sportsmen of different age at a "stamina run". All sportsmen started at the same moment and were supposed to jog on a relatively slow speed as long as they could. The medical team recorded the time when each person stopped from exhaustion. The records are in the spreadsheet table.

SHEE	A	B	C	D
1	Age	Time		
2	34	17.5		
3	21	20		
4	27	19.8		
5	18	22.9		
6	42	14.5		
7	33	16		
8	28	20		
9	50	12.2		
10	36	18.3		
11	44	13.2		

(a) (Refer to Measurements) Draw a scattered graph based on the data. Make the calculation of parameters of the regression line and draw it.

EXERCISES C

Exercise 1.
What time can we expect from a person 40 years old?

Exercise 2.
What approximate age corresponds to 19-minutes run capability?

Exercise 3.
Why is the regression line declining?

SOLUTIONS to EXERCISES C

Exercise 1.
The parameter *a* is approximately equal to -0.32; *b* is approximately equal to 28.15:

$$y = -0.32x + 28.15$$
$$y = 15.35$$

The 40-year old person can run about 15 minutes.

Exercise 2.

$$a = -0.32$$
$$b = 28.15$$

$$19 = -0.32x + 28.15$$
$$x = 28.125$$

The person capable to run 19 minutes is about 28 years old.

Exercise 3.
The older runners have not so big stamina. The duration of their run declines with their age. So does the regression line.

REFERENCE

[LM] Douglas A. Lind and Robert D. Mason, *Basic Statistics for Business and Economics*, Irwin/McGraw-Hill, 1997. ISBN 0-256-19408-4

