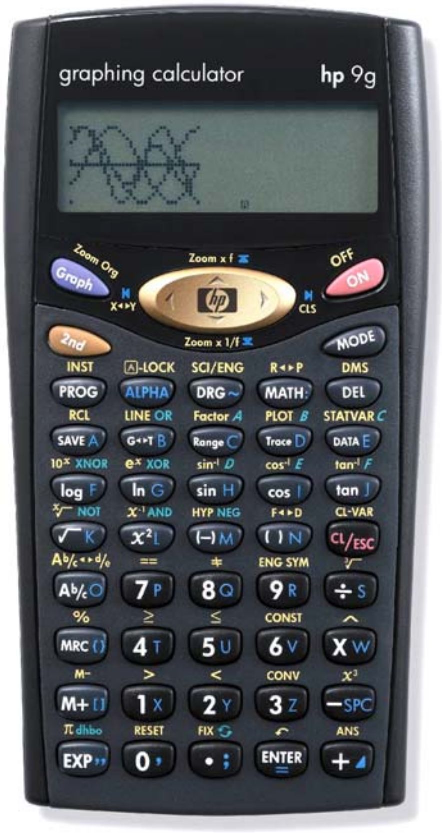
hp calculators

HP 9g Statistics – Linear Regression

Linear Regression

Practice Solving Linear Regression Problems



Linear regression

A regression of *y* on *x* is a way of predicting values of *y* when values of *x* are given. If the regression is based on a straight line graph, it is called a linear regression, and the straight line is called the regression line.

The regression line (sometimes referred to as the line of best fit) of *y* on *x* is then the line that gives the best prediction of values of *y* from those of *x*, and is:

$$\boxed{y = a + bx}$$

$$\text{where} \quad a = \frac{\sum y_i - b\sum x_i}{n} \quad \text{and} \quad b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

n being the number of data pairs. (Note that the regression line of *x* on *y,* which is usually different from the regression line of *y* on *x,* can be found by interchanging *x* and *y* in the above expressions). *a* and *b* are known as the linear regression coefficients. The independent variable is the regressor, and the dependent variables is called regressand. The coefficients are found by minimizing the sum of the squares of the vertical distances of the points from the line (i.e. the sum of the squares of the residuals). This method is known as least squares.

The correlation coefficient is a measure of the amount of agreement between the *x* and *y* variables, and is given by:

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

When r is positive, the correlation is positive, which means that high values of one variable correspond to high values of the other. Conversely, if r is negative then the correlation is negative: low values of one variable correspond to high values of the other. An important property of r is that $-1 \le r \le 1$. The ±1 values correspond to a perfect correlation: real values and estimates are exactly the same. If r = 0 then there's no correlation: x and y are uncorrelated.

On the HP 9g, linear regressions are calculated in the STAT operating mode. First of all, press MODE 1 X to display the STAT menu, and select D-CL and press ENTER to clear previous data. Then, press MODE 1 X again and select REG, and press ENTER . The REG menu is now displayed showing the types of regression available. Select LIN and then press ENTER . You're now ready to carry out regression calculations on your calculator, which are illustrated in the following examples.

Practice solving linear regression problems

Example 1: A quality control engineer notes a relationship between the amount of chemical added to a batch, and the final concentration of the chemical in the final product. The following table shows the weight in grams added (*x*) and the weight in the final product (*y*):

x	2	1	6	3	7	6	9
y	3	1	5	5	7	8	8.5

Find the linear regression line and the correlation coefficient for this data.

Solution: First of all, we have to enter the given data, which is really easy on the HP 9g. Just press DATA E , select DATA-INPUT in the displayed menu and press ENTER . "x_1=" appears in the entry line, and below is displayed the word LIN (for linear regression). We can now enter the data by pressing:

2Y ⌄ 3Z ⌄ 1X ⌄ 1X ⌄ 6V ⌄ 5U ⌄ 3Z ⌄ 5U ⌄ 7P ⌄ 7P ⌄ 6V ⌄ 8Q ⌄ 9R ⌄
8Q • ; 5U ⌄

Data pairs are entered in order, the x values first. Notice that values are actually entered by he ⌄ key, not the ENTER key. This is because you may wish to calculate a value first (e.g. A ENTER + B % ANS ENTER) , keep in mind that the history stack (i.e. a log of past calculations) is not available in STAT mode, but the ANS variable can be used, and the most recent calculation can be retrieved to be edited. Correcting data is as easy as using the ⌃ and ⌄ to display the wrong value and change it. The new value replaces the old one, you don't need to press CL/ESC to remove it from the entry line.

Once all the values have been entered, results can be displayed by pressing 2nd STATVAR C. This displays a menu with five options. Select the desired option using t the ‹ and the › keys. Their values appear in the result line. Pressing ENTER puts the selected variable (its name, not its value)) into the entry line for further calculations. This menu works much like the CONST menu.

Answer: Expressed to two decimal digits, a = 1.22 and b = 0.85, therefore the regression line is: $y = 1.22 + 0.85x$. The correlation coefficient is 0.91, which means that the correlation is positive and that it is quite a good fit since r is close to 1. However, exactly how far away from this value the correlation can be and the equation still be considered a good predictor is certainly a matter of debate.

Example 2: Plot the regression line of the previous example on your HP 9g.

Solution: The STATVAR menu being still displayed, to plot the regression line just press the Graph key. As simple as that.

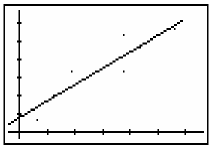


Figure 1

Answer: Figure 1 shows the resulting display. Note that data points are also plotted.

Example 3: If the engineer adds 4 grams of the chemical, what will be the concentration in the final product?

Solution: Predicted values can be easily calculated using the regression line, but the quickest way is to use the STATVAR menu again. Press 2nd STATVAR C , select y' and press ENTER . The entry line now reads $y'()$ with the blinking cursor after the left parenthesis. y' returns a predicted y value given an x value, that is, returns $1.22 + 0.85x$ (remember that these numbers are shown to two decimal digits in this document, but not on your calculator). Enter the given x value: 4T and press ENTER to calculate the predicted concentration.

Answer: $y'(4) = 4.63$. In many textbooks y' is written as $\hat{y}$. It is important to understand that the actual concentration may be different.: the regression line is just a mathematical model of the reality.

Example 4: In order to obtain a concentration of 10.5, how much chemical should she add?

Solution: Once again press 2nd STATVAR C but select x' this time, and press ENTER to put this function into the entry line. x' returns $\frac{y-a}{b}$. Enter the given y value: 1X 0, •; 5U and press ENTER to calculate the estimated amount of chemical.

Answer: $x'(10.5) = \hat{x}(10.5) = 10.89$ grams.

Example 5: The previous examples are based on the regression of the final concentration (*y*) on the amount of chemical added (*x*). Would the last result obtained be equal to $y'(10.5)$ if we were studying the regression of *x* on *y*?

Solution: The most likely answer is no. *y* is the *dependent* variable and *x* is the *independent* variable. Their roles cannot be interchanged. If we interchange x and y, we change our experiment, the results of which may well be meaningless. Let's see what happens, by swapping the given data, and then we'll find $y'(10.5)$. Press DATA E and reenter the data in the following order:

3Z ▼ 2Y ▼ ▼ ▼ 5U ▼ 6V ▼ 5U ▼ 3Z ▼ ▼ ▼ 8Q ▼ 6V ▼ 8Q •; 5U ▼ 9R ▼

To find $y'(10.5)$ press:

2nd STATVAR C select y' ENTER 1X 0, •; 5U ENTER

Answer: According to the new regression, the predicted value is now 9.88 grams. The regression line is now $x = -0.38 + 0.98y$ (where x is still the amount of chemical added and y is the concentration), which is *not* the same as before ($x = -1.44 + 1.18y$)

Example 6: By polling fifty people, a survey taker obtained the following data:

$$\sum x_i = 3333,\ \sum y_i = 459.9,\ \sum x_i^{\ 2} = 231933,\ \sum y_i^{\ 2} = 4308.57 \text{ and } \sum x_i y_i = 30549.75$$

Judging by the correlation coefficient , is there a linear relation between *x* and *y*?

Solution: *r* can be calculated by the formula given on page 2:

$$r = \frac{30549.75 - \frac{3333 \cdot 459.9}{50}}{\sqrt{\left(231933 - \frac{3333^2}{50}\right)\left(4308.57 - \frac{459.9^2}{50}\right)}}$$

Let’s enter it into the entry line by pressing:

( )N 3Z 0, 5U 4T 9R •; 7P 5U −SPC 3Z 3Z 3Z 3Z XW 4T 5U 9R •; 9R ÷S 5U 0, › ÷S
√K ( )N 2Y 3Z 1X 9R 3Z 3Z −SPC 3Z 3Z 3Z 3Z x^2L ÷S 5U 0, › ( )N 4T 3Z 0, 8Q •; 5U
7P −SPC 4T 5U 9R •; 9R x^2L ÷S 5U 0, and finally ENTER .

Answer: r = –0.12, so we can assume there’s no linear relation at all.

Example 7: An experimenter obtained the following data:

x	300	420	450	500	610	780	800
y	11.1	12.2	12.5	13	15.6	15.8	16.1

Plot the data on your HP 9g and find out if there is a linear relation between *x* and *y*.

Solution: First of all, let’s clear any previous data: MODE 1X , select D-CL and press ENTER . This is not required in this example: since N did not change, new data will overwrite the old ones. But, it’s a good habit to clear previous data before starting a new regression calculation.

Scatter graphs are drawn in the 2-VAR mode, so press: MODE 1X , select 2-VAR and press ENTER .Let’s now enter the data: DATA E , select DATA-INPUT and ENTER , then press:

3Z 0, 0, ⌄ 1X 1X •; 1X ⌄ 4T 2Y 0, ⌄ 1X 2Y •; 2Y ⌄ 4T 5U 0, ⌄
1X 2Y •; 5U ⌄ 5U 0, 0, ⌄ 1X 3Z ⌄ 6V 1X 0, ⌄ 1X 5U •; 6V ⌄
7P 8Q 0, ⌄ 1X 5U •; 8Q ⌄ 8Q 0, 0, ⌄ 1X 6V •; 1X ⌄

To plot the graph press: 2nd STATVAR C Graph . Figure 2 shows the resulting graph. We can clearly see that point (610, 15.6) is anomalous, and therefore we will remove it from the data set. To do so, press:

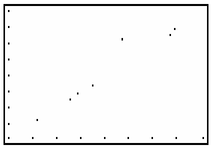


Figure 2

DATA E , select DATA-INPUT, ENTER , ⌃ seven times, and DEL .

Let’s now find the linear correlation coefficient. We have to set the proper STAT mode first: MODE 1X , select REG, press ENTER , select LIN and press ENTER again. Regression variables are displayed in the 2nd STATVAR C menu.

Answer: r = 0.9997, so there’s strong evidence that the relation is linear. The regression line is

$$y = 8.03 + 0.01x$$