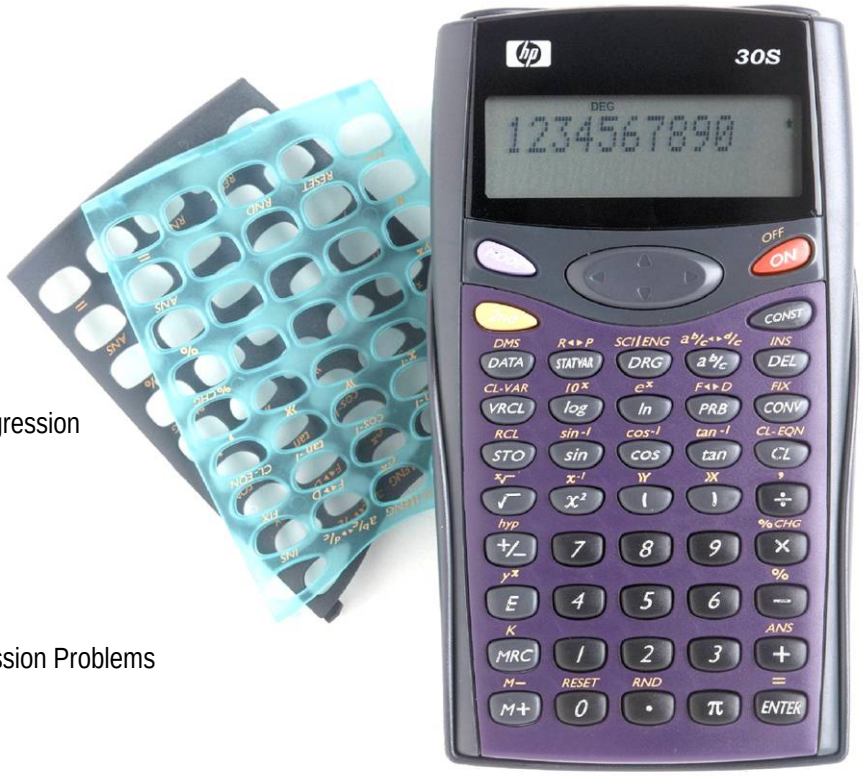
hp calculators



HP 30S 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:

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

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

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:

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

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 30S, linear regressions are calculated in 2-VAR STAT operating mode. First of all, let's clear any previous data. To do so, press MODE 7 to display the STAT menu, and then select CLR-DATA using the ◄ and ► keys, finally press ENTER to confirm. Next, press MODE 7 again, select 2-VAR, and press ENTER . You're now ready to carry out regression calculations on your calculator, which are illustrated by 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 30S. Simply pressing DATA makes your calculator prompt you for the first value displaying "x_1=" in the entry line. Let's now enter the data by pressing:

2 ▼ 3 ▼ 1 ▼ 1 ▼ 6 ▼ 5 ▼ 3 ▼ 5 ▼ 7 ▼ 7 ▼ 6 ▼ 8 ▼ 9 ▼ 8 . 5 ▼

Data pairs are entered in order, the x values first. Notice that values are actually entered by pressing the ▼ 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 ▼ keys to display the wrong value and change it. The new value replaces the old one: you need not press CL to clear the entry line.

Once all the values have been entered, results can be displayed by pressing STATVAR . This displays a menu with seventeen variables. Select the desired result using the ◄ and the ► keys. 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: 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 STATVAR , select y' and press ENTER . The entry line now reads $y'()$ with the blinking cursor placed on the right 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: 4 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 well be different.: the regression line is just a mathematical model of the reality.

Example 3: In order to obtain a concentration of 10.5, how much chemical should she add?

Solution: Once again press STATVAR 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: 1 0 . 5 and press ENTER to calculate the estimated amount of chemical.

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

Example 4: 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 and reenter the data in the following order:

3 ▼ 2 ▼ ▼ ▼ 5 ▼ 6 ▼ 5 ▼ 3 ▼ ▼ ▼ 8 ▼ 6 ▼ 8 . 5 ▼ 9 ▼

To find $y'(10.5)$ press:

STAT VAR ► ► ► ► ► ► ► ► (to select y') ENTER 1 0 . 5 ENTER

Answer: According to the new regression, the predicted value is 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 5: 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 using the formula given on page 2:

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

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

(3 0 5 4 9 . 7 5 − 3 3 3 × 4 5 9 . 9 ÷ 5 0)
÷ √ (2 3 1 9 3 3 − 3 3 3 3 x² ÷ 5 0) (4 3
0 8 . 5 7 − 4 5 9 . 9 x² ÷ 5 0 and finally ENTER .

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

Example 6: 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

Determine whether there is a linear relation between *x* and *y*.

Solution: First of all, let's clear any previous data: MODE / , select CLR-DATA and press ENTER . This is not required in this example: since the number of data items did not change, new data would overwrite the old ones. But, it's a good habit to clear previous data before starting a new regression calculation. 2-VAR mode was already set, so we can now enter the data as follows:

DATA 300 ▼ / / . / ▼ 420 ▼ / 2 . 2 ▼ 450 ▼ / 2 . 5 ▼ 500 ▼ / 3 ▼ 6 / 0 ▼ / 5 . 6 ▼ 780 ▼ / 5 . 8 ▼ 800 ▼ / 6 . / ▼

Let's now find the linear correlation coefficient: STATVAR ◀◀◀◀◀◀ . Rounded to four decimal places, r = 0.9624. Even though it is quite close to one, the experimenter expected a more conclusive result. By plotting a scatter graph (figure 1), she notices that point (610, 15.6) is anomalous, and is consequently removed from the data set. To do so, press:

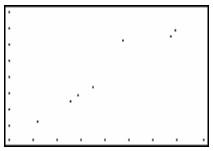


Figure 1

DATA ▲ seven times, and DEL (NB: *not* CL).

The new correlation coefficient is displayed as above. i.e. by pressing STATVAR and then the left arrow key six times.

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

$$y = 8.03 + 0.01x$$

Example 7: Find the power curve $y = m \cdot x^n$ that best fits the following data:

x	0.50	0.75	1.00	1.25	1.50	2.00
y	0.47	1.43	3.15	5.75	9.45	20.68

Solution: This problem can be solved on your HP 30S by making the substitutions y' = ln y, and x' = ln x. The model becomes: $y' = \ln m + n \cdot x'$, which is a linear form. Clear the statistical data (MODE / , select CLR-DATA and press ENTER) and enter the new data as follows:

DATA ln . 5 ▼ ln . 4 7 ▼ ln . 7 5 ▼ ln / . 4 3 ▼ ln / ▼ ln 3 . / 5 ▼ ln / . 2 5 ▼ ln 5 . 7 5 ▼ ln / . 5 ▼ ln 9 . 4 5 ▼ ln 2 ▼ ln 2 0 . 6 8 ▼

In the STATVAR menu (STATVAR), we find that a = 1.140696782 and b = 2.728608754. Since b = n and $a = \ln m$, then n = 2.728608754, and $m = e^a$, which can be calculated as follows:

CL 2nd e^x STATVAR ◀ (eight times) ENTER ENTER

Answer: Rounding to two decimal digits, $y = 3.13 \cdot x^{2.73}$