Performing Linear Regressions With 1-2-3
(PC Magazine Vol 3 No 20 Oct 16, 1984 by P. Jeanty)

Linear regressions are equations that estimate the degree of linear relationship between two sets of variables. They also indicate the equation of the line along which the variables are related. If the variables are sufficiently related, the linear equation produced by the regression can be used to predict the probable value of the dependent variable based on the known value of the independent variable.

Suppose that you were interested in the relationship between the lengths of two bones in fetuses: the femur (thigh bone) and the humerus (arm bone). Using 200 pairs of data for femur and humerus length, one can produce a graph representing the femur (the first variable) on one axis and the humerus (the second variable) on the other axis. 123 can produce this kind of "scatter diagram" directly.

But while scatter diagrams help demonstrate that there is a relationship between the size of the femur and the size of the humerus in fetuses, they can't provide a complete understanding of that relationship; they do not allow you to look at the size of the femur and predict the size of the corresponding humerus. To do that, you need an equation that describes the relationship between the two variables mathematically. The statistical procedure that produces such an equation is called "curve fitting." Linear regressions perform the simplest kind of curve fitting: they fit straight lines. The general equation for a straight line is: $y=a+bx$ where y is the predicted variable (the humerus in our example), x is the observed variable (the femur), and a and b are the two coefficients to be discovered by the linear regression.

The first step is to introduce the data. Create a clean worksheet by using the Worksheet Erase Yes command sequence (/WEY). Place the title "FE" (for femur) in the A1 cell and set the column width to four characters with the command /WCS4. Similarly, enter "HU" (for humerus) in the B1 column, again using /WCS4 to set the width. To separate the titles from the data, go to A2 and enter \=. You extend the data separator across the worksheet with the Copy command (/C) and specify A2 as the range From, and A2..N2 as the range To. (See LINREG.WKS.)

Now you're ready to enter the data pairs into columns, beginning with A3 and B3. Keying in 198 pairs of femur and humerus lengths is a task, but it must be done if you want to duplicate the figures calculated in this article.

After entering the data pairs, the next step is to go to cell G3 and type = n, followed (in G4 through G20) by the series of equates shown in the Figure. These identify upcoming calculations in the corresponding F3 through F20 cells, but additional formulas must first be entered, or cells will begin to fill with error messages.

As you noticed while entering the equations, the calculations require the sums of x^2 and y^2, and the product of x and y. Columns L, M, and N will be used to hold these calculations. In cell L1, enter the idenfifying label x^2; put y^2 in cell M1; and type x*y in N1. Dropping down to Row 3, enter +A3^2 in cell L3, +B3^2 in M3, and +A3*B3 in N3. At this point, you can use the Copy command to fill in all the calculations for each column for as many data pairs (rows) as you entered. With the highlight on L3, type /C and enter L3 in response to the From request, and L3..L200 in response to the To request. It will take only a few seconds for 123 to calculate the results. Repeat this procedure for the M and N columns. (Some 123 users may choose to use the Range Name Create [/RNC] command instead of the Copy command, but the Copy procedure is perfectly adequate for a small database such as this.)

To graph the scatter diagram of column B onto Column A, select the XY option after typing /GT (the Graph Type command). Select the x-axis (horizontal) for the first variable with the /GX command. The /GB command places the second variable on the y-axis. Entering /GOF keeps 123 from tracing lines between your data points, and the program displays the menu Graph A B C D E F. Point the cursor to B (the best symbol to use) and hit Enter. Another menu appears offering you the choice of Line, Symbol, Both or Neither, and you simply select Symbol and hit Enter. Now type /GV to view the graph.

You can give your graph a title (LINEAR REGRESSION) with the /GOTF command sequence. With /GOTS you can add a second title (enter/F3, to indicate the number of cases in your example). A title for the x-axis is introduced in the same way: /GOTX and \AL; the y-axis is identified with /GOTY and \BL. When you use this worksheet for computing linear regressions with other parameters, the titles will automatically be update along with the graphs.

The time has come to proceed to the linear regression by filling in the missing rows in column F. The following indicates what should be entered (with the Lotus calculating function-sign @, as shown) for each of the F cells:

Cell	Entry
F3	@COUNT(A3..A200)
F4	@SUM(A3..A200)
F5	@SUM(B3..B200)
F6	@SUM(L3..L200)
F7	@SUM(M3..M200)
F8	@SUM(N3..N200)
F9	@AVG(A3..A200)
F10	@AVG(B3..B200)
F11	+F6-F4^2/F3
F12	+F7-F5^2/F3
F13	+F8-F4*F5/F3
F15	+F13/F11
F16	+F10-F15*F9
F17	+F13/@SQRT(F11*F12)

When entering this information, remember to use the plus sign (+) where indicated, and do not put any spaces between the items on which calculations are being made.

When you have made these calculations, you will have arrived at the equation for the line describing the relationship between femur length and humerus length: y=F16+F15*x. You can use this equation to go back and predict the value of y for each x value. You can then compare the predicted values with the observed values.

To do this, enter ^PredictedY as a label in C1, then drop down to C3 and enter +F16+F15*A3. 123 answers with 12. While it might be tempting to use the same copy procedure used earlier for columns M, N, and L to fill in all the values of C, one procedural change must be made. You want the addresses F15 and F16 to be absolute rather than relative, so go back to C3 and enter the editing mode, function key F2. Change the formula to:

```
+$F$16+$F$15*A3
```

The /C can now copy this formula C3..C200.

You can now update the graph with the /GA command by typing the range C3..C200. To distinguish the new dots from the original data, enter the /GOFA command, point the cursor at the Line option in the menu, and Enter. The predicted value (y) will be represented by a continuous line without associated symbols. Type QV to view the

updated graph.

You can now predict the value of the second variable based on a given value of the first variable. The next question is, how close to this predicted value can you expect the observed value to fall?

The accuracy of the prediction is determined by the coefficient of correlation, held in cell F17. This coefficient, r, measures the strength of the relationship between the regression line and the data pairs. The closer r is to one, the more closely the data pairs will tend to conform to the regression line. The closer r comes to zero, the more the plotted data will resemble an amorphous cloud. To test this, try changing the data in column A or B by inserting a value two or three times larger than the current one. The newly created dot will fall far outside the range of other dots on the graph. Hit the F10 key and you will see the value in cell F17 decrease.

Now you're ready to deal with the question of confidence limits. They control the certainty with which you can say that the observed value will fall within a given range of the predicted mean value (produced in column C). The size of this range is expressed in standard deviations, so you must first compute the standard deviation of the points around the regression line. Start by defining cell F20:

```
F20 is @SQRT((1/(F3-2))*(F12-(F13^2/F11)))
```

Obviously, the size of this range will determine how likely the observed value is to fall within it. For the data considered here, a range of plus or minus 1.66 standard deviations would encompass 90% of the observed values, leaving a 10 percent chance that a correct value might fall outside the acceptable range. If you can live with that large of a margin of error, use the 1.66 factor in the next calculation. If you desire a stricter standard, leaving only 5% of correct observations outside the range, the size of the range would have to be increased to plus or minus 1.98 standard deviations. You can have more confidence in the wide range, but it is less precise.

With cell F20 defined, go to cell D1 and type the label 2.5, which represents the 2.5th percentile. Next, go into cell D3 and type this formula:

```
+C3-1.98*(@SQRT($F$20^2*(1+1/$F$3+(A3-$F$9)^2/$F$11)))
```

Be sure to include the correct number of parentheses or 123 will beep at you. Again, note use of the dollar sign ($) to indicate absolute reference. Next, copy D3 into D3..D200 to return to the lower percentile.

Repeat the process and define cell E1 as 97.5. The formula to be introduced in E3 will be the same as the one in D3, except that the first minus sign is replaced with a plus sign:

```
+C3+1.98*(@SQRT($F$20^2*(1+1/$F$3+(A3-$F$9)^2/$F$11)))
```

Do not copy the contents of cell D3 into E3. That would change the value of relative cells such as C3 and A3 into D3 and B3! Retype the formula as indicated. When this is done, copy the contents of E3 into E3..E200.

That's it. To put some icing on the cake, update the graph by including the two new columns, D and E, with the procedure used for column C. This time add the new ranges in C and D as well as the /G prompt.

Rather than junk the worksheet now, save it again under a different name. Error messages will appear in every cell that held a computed value, but that's all right. Save this "template" under the LIN_REGR name. The next time you need to compute a linear regression, just define the numbers that you want to calculate in your worksheet, give them a name, save the worksheet, call the LIN_REGR worksheet, issue the /FCCN command, and answer 123's question of which names in which worksheets you want to combine. Using the template avoids having

to export data to statistical programs and cuts the time to obtain the answer.

You can use the template with just about any data you desire, even the thickness of PC Magazine agains the number of the issue. Just remember, predictions are reliable only with a high coefficient of correlation r; it is very dangerous to use this curve to predict events outside the range of observed values. If you apply the equation of your son's growth against his age, it may predict that by age 70 he will be 10 feet tall.

##
##
##