Massachusetts Institute of Technology
Physics Department

8.01X **Fall 2000**

Linear Regression on Xess

The default version of Xess (4.1) on Athena works fine, but has the same drawback as the previous versions: Until you get used to it, the most valuable (to us) applications of Xess are hard to use. The on-line Help feature, for instance, is quite extensive, but it's in hypertext with very few links or anchors. This may be addressed soon, we hope. Online help from Athena is available at **Xess Help on Athena** (linked from the **8.01X** pages).

These notes do not give detailed instructions, but rather refer to an example that may be used as a template for simple linear regression. You almost certainly have statistics capabilities on your handheld calculator, and these notes will show how Xess improves on this. One immediate advantage is that all of the data are displayed simultaneously, a great advantage for changing data entries. The example also shows how the data may be graphed and displayed in a variety of forms.

To start up Xess, either start from the Dash under Numerical/Math → Spreadsheets, or, at the Athena prompt, do

```
athena% add xess
athena% xess&
```

and wait a bit; Xess is a large program and it takes a while.

For now, perhaps the best thing to do would be to copy the template I've made, or load it directly using the "Open" command in the "File" menu, or Control-o. It's in

```
/mit/8.01x/www/other/lnrg.xs4
```

(or download from the **8.01X Xess Intro** page) and it's pretty simplistic. If you copy this file, you must include the ".xs4" in the filename (the suffix is added by default if you download the spreadsheet).

When the Xess worksheet appears, you will see that the program is pretty much menu-driven. You are of course welcome to explore the different features. For now, I'm going to leave things like fonts and sizes pretty much the way they are, and I won't mess with labels and titles any more than is needed to save the work.

These notes do not give detailed instructions, but rather refer to Problems 29-31, Page 696 in the current **18.01** text, *Calculus with Analytic Geometry* by

George F. Simmons, Second Edition. When you have done this example, your work may be used as a template for other simple linear regressions. Or, you may be inclined to investigate further, more involved uses of Xess.

You almost certainly have statistics capabilities on your handheld calculator, and these notes will show how Xess improves on this. One immediate advantage is that all of the data are displayed simultaneously, a great advantage for changing data entries. The example also shows how the data may be graphed and displayed in a variety of forms.

The solution to Problem 29 of the above set is

$$m = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}, \qquad b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}.$$

In the above, each sum is assumed to be over the range $i = 1 \ldots n$. Note the distinction between the sum of the squares and the square of the sum; that is,

$$\begin{aligned} \sum x_i^2 &= x_1^2 + x_2^2 + \ \ldots + x_{n-1}^2 + x_n^2 \\ \left(\sum x_i\right)^2 &= \left(x_1 + x_2 + \ \ldots + x_{n-1} + x_n\right)^2, \end{aligned}$$

and these two quantities will not be the same, of course. In fact, Problem 31 on Page 696 shows that they cannot be the same, and that the denominator in the expressions for both m and b will not vanish.

One of the main points of using linear regression is that the needed sums can be changed quite easily as data are entered or altered. In fact, each sum may be represented as the dot product of two vectors, each with n elements. This sort of matrix manipulation is what calculators and computers do with ease. In fact, the equivalent calculations in MATLAB (which is an acronym for “MATrix LABoratory”) use, quite naturally, matrix formalism to both enter the data and do the algebra to obtain the above forms for m and b.

What we will do is enter the values of the independent variable (we’re calling these values x_i) in column A and the values of the dependent variable (y_i) in column B. To enter data, just click on the cell and type the numbers. Either a return or a move with the arrow keys will enter the value into the worksheet, as you can see. Whenever you type a quantity that is purely numeric, Xess will interpret it as a number (thus allowing calculations). This is good.

The data from Problem 30 on Page 696 have been entered, and so the integers 1-4 are in column A and the corresponding y-values, in this case 1.7, 1.8, 2.3 and 3.2, in column B. By now, you might realize that we’re dealing with a fairly simple case, with only four data pairs. Xess could handle several thousand, if necessary. That is, please don't judge Xess on the basis of the simple numbers we're using.

Now, click in cell D1 (there's a reason; we'll come back to column C in a bit). You should see in the edit window the string (I'll use teletype font, just to distinguish from the body of these notes, and the commands are not case-sensitive):

```
=@LINCOEF(A1..A4,B1..B4)
```

and enter to see the values of what we would call m and b in the respective cells D1 and D2.

At this point, if you want to see the troubles involved in using the online help, call up the Help Index from the upper right corner. The stinker here is that what we want is *not* under "Statistics", but is considered one of the "Embedded Tools". You may find that to use the online help, you pretty much have to know where to look ahead of time.

The fact that the numbers came out so nice is that they were rigged for easy checking. In general, your data will never allow neat numbers like this.

While we're here, click in cell D3 to show the string

```
=@CORR(A1..A4,B1..B4)
```

which generates the "correlation coefficient", a measure of how good a fit the line is to the data. A correlation coefficient close to unity indicates that the linear fit is good. Correlation coefficients are not addressed in Simmons; they're not tough, but they need some introductory probability theory, and there isn't enough room in the text or time in the term for that much fun. The expression for this coefficient, conventionally denoted as r, is

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

where each sum is from $i = 1 \ldots n$. By the way, if you've done Problem 31 on Page 696, you know that the argument of each square root cannot be negative, and you should know what it would mean if either were zero.

The number returned in the default format is too precise; to change the precision, click in the cell, go to the "Format" menu, and call up "Cell Format". When this window appears, change the Cell Format to "Fixed", and then choose "2" for the number of decimal places (this has been done in this example template already).

As it turns out, the Correlation Coefficient is indeed one of the "Statistics Functions", not one of the "Embedded Functions". Go figure.

Back to the spreadsheet: Click in cell C1 and observe the string

```
=@LINFIT(A1..A4,B1..B4)
```

This gives the values of $m\,x_i + b$, for the x_i in column A and the values of m and b already found (should still be in cells D1 and D2). The simple numbers allow easy checking of the numbers displayed in column C. The "LINFIT" function is one of the "Embedded Tools".

Now for one of the real advantages of this program: Graphing the data and the best-fit line. With the mouse, highlight the box of cells with corners A1 and B4. Go to the "Graph" menu, and pull down to "New Graph/Scatter Graph". A graph of the data will appear (you may want to resize this thing right away; the default size is rather tiny).

To get a simultaneous plot of the best-fit line, you need to add to the data sets. Do this from the "Edit" menu. When the window appears, it may not seem at all clear what to do (it sure wasn't to me). In the top line, you should see the number "1" at the far left. This means that the first data set plotted, which is the only one so far, is the values in cells B1 to B4 as a function of the values in cells A1 to A4. You'll see that there is "No Line" in this scatter plot, which is why we chose it. If you wish, you can change the color or other features, or add a legend.

On the top line, point and click just to the right of the small rectangles under the "1". This will allow changes in the features for the second data set, which you need to set. In the box for "X Data", type `a1..a4` and for "Y Data" type `c1..c4` (again, the entries are not case-sensitive). Now, however, you do want a line graph, not a scatter graph, so go to "Line Style" and select the style you wish ("Dashed" is recommended, but not strongly). Set the other features as you wish (but don't go overboard with "Fill Style"; that could make the graph look goofy).

An important item that should be done by default but which isn't is to have the graph redraw automatically when data is changed. To do this, in the graph window select "Options/General", and make sure that the "Redraw on Calculation" button is "pushed". While you're here, you can give your graph whatever title you want.

If you want to save the graph, this is of course done by "File/Save As .." from the graph edit window. The name of the graph will not be the same as the title unless you make it so. You need to save the graph if, for instance, you wish to insert the graph into the sheet. To insert, click on an empty cell (E1 comes to mind), select "Graph/Insert in Sheet .." and when the window appears, highlight the name of your graph and "Apply". Neat, isn't it?

If you're having trouble seeing what to do, view the "Demoplot", from the "Graph" menu.

Now for another important feature, hinted at earlier: Change the data. That is, click in any of the eight cells where the data are entered and edit that number (but don't go too wild with 6.023×10^{23}, or its ilk). You'll see that the coefficients and the best-fit line entries change, and the graph redraws. If you change one of the elements in column A, for instance changing the entry in A1 from 1 to -1, even the horizontal range of the graph changes. (If you do this, you'll note what happens for negative numbers.)

Experience suggests that you have already begun to see what else you can do in terms of options and features. That is the idea.