Math 182

Assignment #4: Least Squares

Introduction

In any investigation that involves data collection and analysis, it is often the goal to create a mathematical function that "fits" the data. That is, a function whose values are close to the data values at the corresponding values of the independent variable.

While Maple is able to perform least-squares analysis using a collection of routines in the 'Statistics' package, the point of this assignment is to give you an introduction to the underlying analysis. You will perform a simple least-squares analysis of a small data set. Your investigation will require both pencil-and-paper and Maple-based analysis. Along the way you will learn how to use Maple to solve a system of equations using `fsolve`.

As you progress, you will come to see why the name "least squares" is appropriate.

Procedure

Suppose an experiment was performed that generated the following set of data:

$x = 15, 30, 45, 60, 90, 120, 150$; $y = 0.485, 0.460, 0.435, 0.410, 0.375, 0.340, 0.300$

Suppose that the theory governing the system under investigation predicts that x and y are related by the model $y = Ce^{-kx}$. Your task is to find the values of k and C that yield a curve that "best fits" the data in some sense.

1. Open a blank Maple worksheet and save it with an appropriate filename. Since you will be submitting a printout of this worksheet, it is important to include comments along with any "Maple work" that you do. This will make your worksheet clearer for anyone reading it (or marking it).

2. It will be convenient later to assign the number of data points to a name, say N. For instance, you will need evaluate a sum of terms involving the data points. You can create the sum fron 1 up to N. In this manner, if later you want to use the worksheet with a new data set, you need only change the value of N at the start of the worksheet, rather than hunt down all references to the size of the data set throughout the worksheet. Assign the number of data points to the name N.

3. Assign the data sequences to names so that we can refer to an entire sequences by its name. In Maple, a sequence is simply a comma-delimited collection. Perform the following assignment:

   ```
   x0:=15,30,45,60,90,120,150;
   ```

   Then assign the y-values to `y0`.

4. Some of the analysis you will be performing on the data will be more convenient if the data is also stored as a Maple list. Execute the following command:

   ```
   x1:=[x0];
   ```

   Do the same for the y-values.

   Individual elements in a Maple list can be accessed in the following way: To access the k^{th} member of the list L, execute the command `L[k]`.

   It's always good to ensure that everything is working, so call up the values of the 4^{th} member of `x1` and the 7^{th} member of `y1` and verify that they are the same as the corresponding numbers in the original data sequences.

5. You'll need a graph of the data. Maple is quite happy to plot the data points, provided you create a list of points to plot. Right now, your data is in two separate lists, `x1` and `y1`. You'll need to create a list of ordered pair from these two lists.

   In Maple, sequences enclosed in [ ] are considered to be ordered, while sequences enclosed in { } are considered to be unordered. In other words, Maple will preserve the order in the former, but not necessarily in the latter.

   Since you want to create a list of *ordered pairs*, the ordered pairs must be Maple *lists*. You will create a list whose members are the two-number lists representing the data points: a "list of lists". For example, a data set consisting of the two points (a, b) and (c, d) can be stored in Maple as `[[a,b],[c,d]]`.

   There are only 7 data points in your data set, so it is not too tedious to create this list from scratch. But it seems that Maple should be able to do this for you. After all, you've already got the x-coordinates and y-coordinates stored in separate lists. Further, if your data set were quite large, it would be very a very tedious (and error-prone) task to enter the data by hand yet again. The following command will create your list of ordered pairs and assign it to the variable $P1$. Study the command carefully to see how it works:

   ```
   P1:=[seq([x1[i],y1[i]],i=1..N)];
   ```

   In particular, note that the `seq` command is itself enclosed in square brackets. Investigate the effect of removing those square brackets. Study the output of the command carefully! Replace the square brackets before moving on.

6. Create a plot of the data points. To ensure that you have only the data points and not a jagged line, use the option `style=point`. Also, you can specify the shape of the symbol used to plot the data points. Execute `?plot[options]` to find out how to make the plot symbols appear as boxes. The command should start as `plot(P1,...);`.

7. You should notice that your raw data do not lay along a straight line. The theory predicts and exponential relationship, so this is not unexpected. On the other hand, it is quite difficult to determine if the graph actually *is* an exponential curve and not some other kind of curve. How can you be sure that your data is in agreement with the theory? In particular, is there a procedure that will allow you to determine the values of k and C that will create a curve that fits your data?

   It is possible to determine the model parameter k from the slope of a certain *straight* line, but you must plot your data in a way that produces a straight line. Here's an example of this idea. Suppose theory predicts that two quantities x and y are related as $y = \sqrt{kx}$. If you were to plot the *square* of y versus x, then the graph would be linear, since $y^2 = kx$. A plot of y^2 versus x will be a straight line through the origin with slope k (if the theory is correct). It is important to understand that the claim that $y = \sqrt{kx}$ is equivalent to the claim that $y^2 = kx$ (provided we understand, in this case, that $y > 0$.)

   Consider the model $y = Ce^{-kx}$. In order for k to be obtainable from the slope of a straight line graph, you need to find a way to convert the model relationship between x and y into one in which the right hand side is a first degree polynomial in x. Find a way to do that, then apply the same transformation to the left hand side. In particular, create new lists `X` and `Y` from the old ones (`x1` and `y1`) such that when you plot `Y` versus `X` you get a straight line with positive slope. [HINT: REMEMBER THAT YOU CAN APPLY A FUNCTION TO EACH NUMBER IN A LIST USING THE `map` COMMAND. YOU MAY WANT TO LOOK UP THE HELP FILE FOR `map` TO REFRESH YOUR MEMORY. ALSO, NOTE IN THE EXAMPLE GIVEN IN THE PREVIOUS PARAGRAPH THAT ONLY ONE OF THE VARIABLES NEEDED TO BE 'TRANSFORMED'. THE OTHER WAS LEFT ALONE. DO NOT ASSUME THAT BOTH `X` AND `Y` NEED TO BE DIFFERENT FROM `x1` AND `y1`.]

8. With these new data values defined, create a new list of data points called `P` and plot this list on a new graph. Again, choose a point style plot using boxes for the points.

9. Although this plot does not look much better than the first, there is a theoretical basis for plotting it in this way. The next step is to find the "best" straight line that these points represent. Clearly the line can't go *through* all the data points, but it is possible to find one that fits the data in some optimal way. The goal of this assignment is to find such a way.

   Ideally, your data points lie directly on a line and no further work is necessary. In general, however, the data points and the points on the best fit line are not the same. Some data points lie above the line, and some below. You want a line in which you have the smallest difference between the data points and the line. That is, for every point the difference between it and a straight line is

   $$\delta_i = y_i - (mx_i + b)$$

   where i indicates the i^{th} data point and m and b are the slope and y-coordinate of the y-intercept, respectively, of the best fit line. If you add up all of these differences δ_i, they tend to cancel out since some are positive and some are negative. You can see that it is possible to have a line that differs significantly from the data and yet for which the sum of the differences δ_i is very small. The sum of differences, therefore, should be rejected as an indication of the "goodness" of the fit. What you really needed is a quantity related to the differences, but which doesn't tend to cancel when you add them up. The absolute values of the differences might work - the *sizes* of the differences, regardless of whether they are positive or negative. It turns out that absolute values are inconvenient, so as an alternative you can consider the *squares* of the differences. The squares of the differences are positive and so the "cancellation of differences" in the sum is eliminated. Define

$$\Delta_i = [y_i - (mx_i + b)]^2 .$$

If you add up all of these squared differences between the data and the straight line, you have the function

$$\chi(m, b) = \sum_{i=1}^{N} [y_i - (mx_i + b)]^2 .$$

What you would like to do is minimize the value of these errors with respect to the parameters m and b. In other words, you are seeking the straigh line, as determined by m and b, that yields the minimum value of $\chi(m, b)$. That is why the sum has been interpreted as a function of m and b.

From Math 181, you know that to minimize a function, you need to find the critical numbers of the function and see if a critical number is the location of a minimum of the function. This involves seeing where the derivative of the function is zero (or undefined, provided that the function itself is defined). Here, you have *two* parameters, and so two derivatives must be evaluated. Since both of these derivatives must be zero, you'll wind up with system of two equations to solve.

The two derivatives you must evaluate are $\frac{\partial}{\partial m}\chi$ and $\frac{\partial}{\partial b}\chi$. The strange 'curly' derivative symbols are used whenever a derivative of a function of more than one input variable is evaluated. The derivative of χ with respect to m is evaluated as you normally would, except that the other variable, b, is treated as a constant. Similarly, in the derivative of χ with respect to b, the variable m is treated as a constant.

To make sense of this, imagine you are standing on a hill. You could investigate how the height of the hill changes as you walk to the east or west. You could investigate how the height of the hill changes as you walk north or south. If x is a variable that denotes east-west position and y is a variable that denotes north-south position, then in the first investigation you are examining the height of the hill as x is varied but y is fixed, and vice versa for the second investigation.

Differentiating $\chi(m, b)$ with respect to m and making the result zero leads to the equation

$$\sum_{i=1}^{N} [y_i - (mx_i + b)] (-2x_i) = 0,$$

and doing the same with the other derivative leads to

$$\sum_{i=1}^{N} [y_i - (mx_i + b)] (-2) = 0.$$

Derive these equations by hand and show your work. Remember that the derivative of a sum of functions is just the sum of the derivatives. Also, remember that you are differentiating with respect to the parameters m and b, not x and y.

10. Now that you have expressions for two equations that must be solved, input them into Maple. Call them `eq1` and `eq2`. Note that you can assign the *entire* equation, including the '=0'

part. Be sure to define the equations in terms of `X[i]` and `Y[i]` and your two unknown parameters m and b.

Double check with your instructor that you have the equations entered correctly. If you are not in the lab while you are doing this, you can always email the file to your instructor to check.

11. Once you have entered the equations Maple will simplify them somewhat, giving you the following two equations:

    $eq1 := -1046.863951 + 103500m + 1020b = 0$ and $eq2 := -12.97550853 + 1020m + 14b = 0$

    These are the two equations that must be solved for m and b. It is not difficult to do this by hand, but Maple is very good at eliminating this tedium from your life!

    As an example, consider the following two equations that you can solve in your head:

    $$A - B = 1 \text{ and } A + B = 3$$

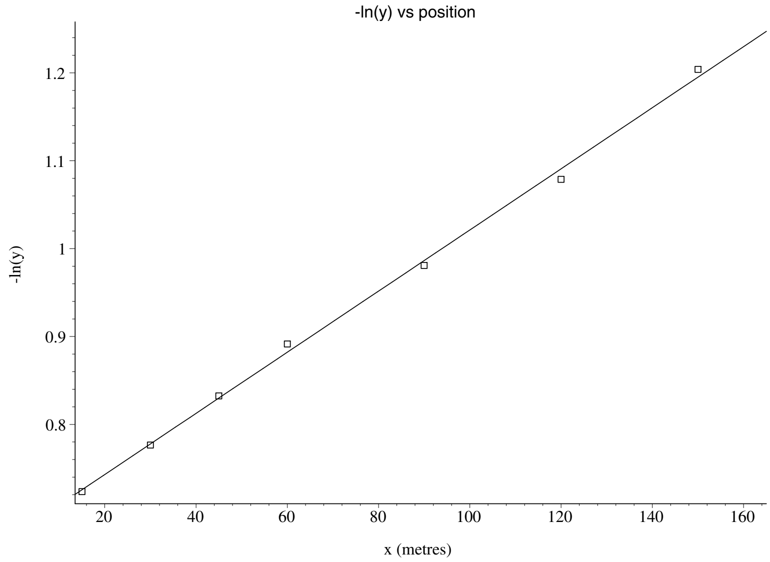
    Using the fsolve command in Maple, you can type `eq:=fsolve({A-B=1,A+B=3},{A,B});`. This not only finds the solutions, but assigns the sequence of solutions to the name `eq`. The solutions it finds should correspond to the ones in your head. Type `A;`. Note that Maple does not give you the value of A indicated in the solution. You must tell Maple to assign the solution values to A and B if you wish to refer to them later. You can do this with the command `assign(eq);`. [In this discussion, the name `eq` can be swapped with any name you wish. You needn't always use the name `eq`.]

    Using this example as a guide, solve the two equations `eq1` and `eq2` and assign the solutions to the parameter names m and b. Note that since you have the equations stored in the names `eq1` and `eq2`, you don't have to type the equations in the braces, just their names.

12. Define the function $f(x) = mx + b$. This is the function whose graph is the best-fit line to your modified data. [As a result of the previous step, m and b have values. You can define the function using m and b, and Maple will fill in the values for you. You needn't cut-and-paste the values.]

13. Plot both the data set P and the best-fit function f in a single graph. A couple of issues arise. First, you want to make sure that the data 'fills' the graph. Maple can help you do this. The commands `min` and `max` will find the minimum and maximum values for a given sequence of numbers. Define `min_x:=min(x0);` and similarly for `max_x`. When you eventually plot the data, you can ensure that the data fills the graph by plotting from a little less than `min_x` to a little greater than `max_x`.

    Second, you want to plot P using the 'point' style and the best fit function f using the 'line' style. Maple will allow you combine plots with different styles, just as it allows you to combine graphs with different colours: use the option `style=[`*style1, style2*`]` where *style1* is either `point` or `line`, depending on whether the P or the $f(x)$ appears first in your function list. Similarly for *style2*. You can also specify the symbol as you've been doing, since Maple knows to apply this option to the graph that has the `point` style. See your instructor if you get stuck. Refer to `?plot[option]` to help you create a final plot which looks something like the following:



14. What are the values of k and C in the equation $y = Ce^{-kx}$?

15. Plot the original data and this equation on the same plot. You have just fit a curve.

16. Why is this method that you have used known as the method of least squares?

For your information...

Maple has built-in procedures to do what you have just done. The procedure is in the 'Statistics' package. In particular, the `Fit` command will do the job. In fact, Maple can fit the original data to the exponential model directly, so that you don't have to modify the data to obtain a linear relationship as you did earlier:

```
with(Statistics):
Fit(C1*exp(-k1*x),x1,y1,x,output=leastsquaresfunction);
fitfunc:=unapply(%,x);
```

Note that in the model, the symbols `C1` and `k1` are used to avoid conflict with the `C` and `k` that may have been assigned values earlier. Also, the `unapply` command takes an expression and turns it into a function.

Compare the values of C and k that you obtained to the values of $C1$ and $k1$ obtained from `Fit`. You can see that they are very close.