Linear Least Squares Approximation Lab

or Fitting a Polynomial Curve to a Set of Data Points.

Part I Introduction

One of the common situations that arise in the real world is as follows. You have a set of data that partially describe a given situation, but you either want to get an educated guess of a future value, or approximate data that lay between the measured data you have. In either case, what you need to do is find a curve (or function) that "best fits" the data. Once you have this function, you can evaluate it at different spots (inputs) and get approximations you are after (outputs).

So, what is meant by the "best fit?" One of the most well known ways is the method called Least Squares. To illustrate, let's start with Figure A, a graph consisting of plotted points (the initial data).

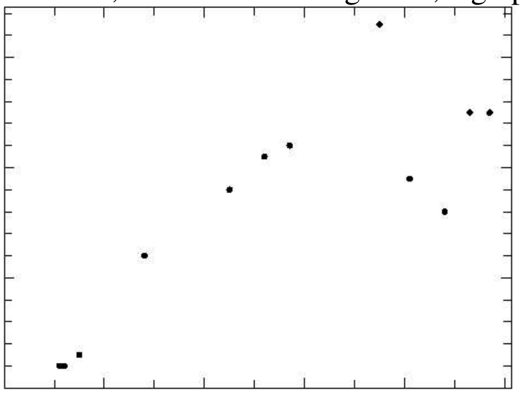


Figure A

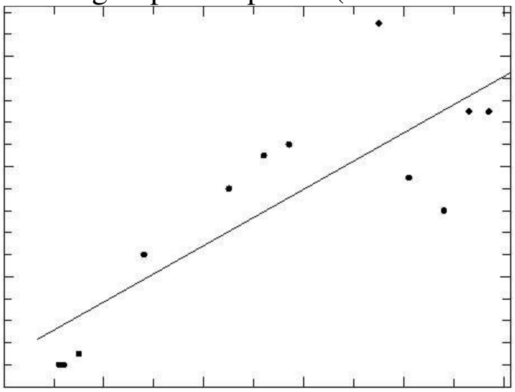


Figure B

If we try to find the best line to fit this data using least squares, we would get the line in Figure B. To get an idea how we would measure how close this line fits, we would look at the vertical error associated with each point, as illustrated in figure C. Now, if we square each error and add up the total square errors, we get the measurement of how well this line fits. Graphically, the square errors are illustrated in figure D.

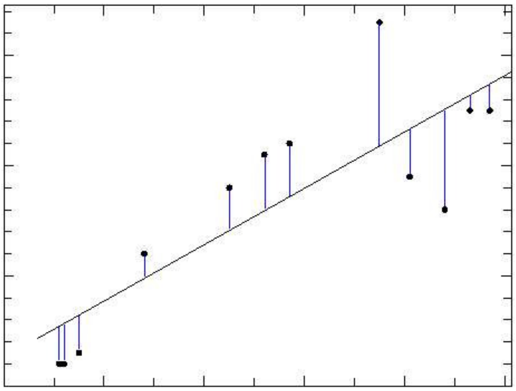


Figure C

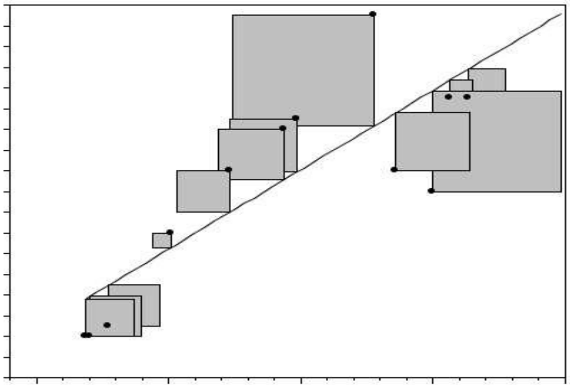


Figure D

Different lines will yield a different sum of the square areas, and so the "best fit" is the one that has the "Least Squares". The reason we use the square of the error is best understood with some knowledge of Calculus, so suffice to say this is how we do it.

Typically, to find a given type of curve that is the least squares solution for a data set, we do need to incorporate the use of tools from calculus. However, if the curves we use to try to best approximate the data are polynomials (linear, quadratic, cubic, ...) instead of more general curves (elliptical, hyperbolic, exponential, logarithmic, etc..) there is a method of computing the least squares solution using matrices. The proof is beyond the scope of this class, so the method will be presented without proof.

We will illustrate this technique with a simple example. Let's start with the following set of data: To find the best linear fit, we use the model $y = mx + b$. Plugging the first point (0,2) into the x and y leads to 2 = b. Point (1,3) leads to 3 = m + b, point (2,6) leads to 6 = 2m + b and point (3,10) leads to 10 = 3m + b. So, we have 4 equations and 2 unknowns (m and b):

x	y
0	2
1	3
2	6
3	10

$$\begin{cases} b = 2 \\ m + b = 3 \\ 2m + b = 6 \\ 3m + b = 10 \end{cases}$$

In matrix form, this would look like

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 6 \\ 10 \end{bmatrix}$$

So we have a system $Ax = b$. The problem with this system is that it is over determined, meaning there are too many equations for the number of variables. Thus, there is unlikely a unique solution. To get the unique solution that is the best fit to the given system, we apply the method of least squares.

The matrix form of this method is as follows: Multiply the equation $Ax = b$ by A^T on the left side. This leads to $A^T Ax = A^T b$, which are called the Normal Equations. This new system has a unique solution, and that solution is the least squares solution.

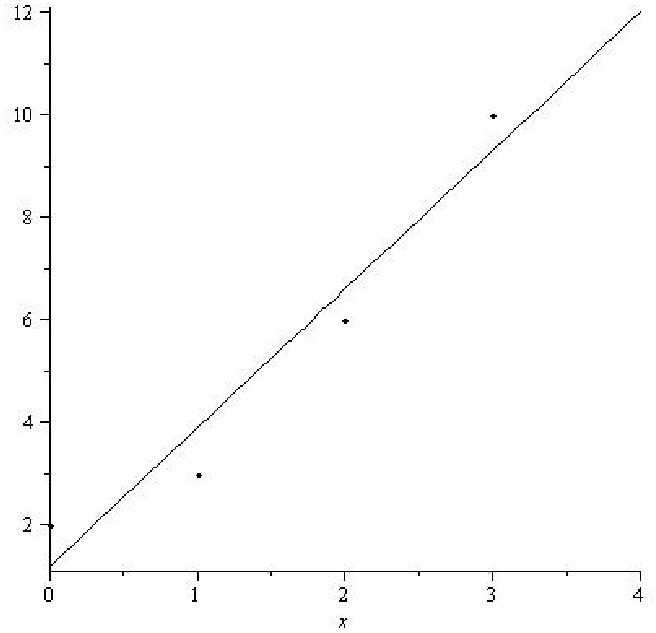
What is A^T ? This is called A transpose and is easy to work out. Basically, transposing a matrix switches the rows with the columns. So, for our $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix}$, $A^T = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix}$.

In our case, the normal equations $A^T Ax = A^T b$ will be

$$\begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 6 \\ 10 \end{bmatrix} \text{ which simplifies to } \begin{bmatrix} 14 & 6 \\ 6 & 4 \end{bmatrix} \begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 45 \\ 21 \end{bmatrix}$$

Solving this system (by elimination, substitution, RREF, Cramer's Rule, inverse matrix) leads to the solution $\begin{bmatrix} m \\ b \end{bmatrix} = \begin{bmatrix} 2.7 \\ 1.2 \end{bmatrix}$, so our "best fit" least squares linear approximation is $\boxed{y = 2.7x + 1.2}$

The following graph includes the 4 points and our linear approximation.



Now it is your turn.

Part II **Vehicle Crashes**

The following facts are based on analysis of data from the U.S. Department of Transportation's Fatality Analysis Reporting System (FARS).

Motor vehicle crash deaths per 100,000 people, 1975-2005:

Year	Rate
1975	20.6
1980	22.5
1985	18.4
1990	17.9
1995	15.9
2000	14.9
2005	14.7

From this graph, we can see that out of 100,000 people, 20.6 people died in a motor vehicle crash in the year 1975.

Use the method of least squares to find the best linear fit $y = mx + b$.

a) Write your system $Ax = b$
b) Write what you have for A^T
c) Write your simplified normal equations ($A^T Ax = A^T b$)
d) Write your least squares solution. (RREF is a good general method for this lab)
e) Graph the points and the linear approximation on the same coordinate axes. To graph your data and the approximate solution on the same coordinate axes, you can use maple, excel, virtual TI and a screen capture program or by hand with graphing paper. See last pages for details. If you find other methods, please share.
f) Use your approximation to estimate the number of deaths per 100,000 you expect in the year 2010.

Part III **Stopping Distance**

When driving, the distance required for a car to come to a complete stop is dependent upon the speed of the car at the time the driver makes the decision to stop. The distance to come to a complete stop has two components, the reaction distance and the breaking distance. The reaction distance is the distance the car travels from the time the decision to stop is made until the moment the break is applied. The breaking distance is the distance the car travels from the moment the brake is applied to the time the car comes to a complete stop. Together the reaction distance and the breaking distance make up the stopping distance of the vehicle.

The following table lists some example stopping distances for different speeds:

Speed (mph)	Reaction Distance (ft)	Breaking Distance (ft)	Total Stopping Distance (ft)
20	22	25	47
30	40	43	83
40	48	97	145
50	55	188	243
60	66	300	366
70	74	455	529

Average total stopping distance of cars on dry, level pavement.

With the data from the Speed column and the Total Stopping Distance column, use the method of least squares to find the best quadratic fit $y = ax^2 + bx + c$.

a) Write your system $Ax = b$ (A should have 3 columns)
b) Write what you have for A^T
c) Write your simplified normal equations ($A^T Ax = A^T b$)
d) Write your least squares solution.
e) Graph the points and the quadratic approximation on the same coordinate axes.
f) Use your approximation to estimate how fast a driver would be traveling if their skid marks were 722 feet long.

Part IV **LINEAR or QUADRATIC? Golf Ball Distance**

When released on a hill with a constant slope, a golf ball falls at a speed determined by gravity, the slope, and friction from the grass. The below data table shows you the results of a specific experiment concerning how long it took in seconds for a golf ball to travel the given distance, in feet. What you need to do is to determine from the data whether you feel a linear model or a quadratic model would be a better fit. You can do this by finding both, or by sketching the data and doing a visual estimation. Once you determine which one you prefer, find that model and graph the resulting curve, along with the given data.

a) Determine whether you will use $y = mx + b$ or $y = ax^2 + bx + c$.
b) Write your system $Ax = b$
c) Write what you have for A^T
d) Write your simplified normal equations ($A^T Ax = A^T b$)
e) Write your least squares solution.
f) Graph the points and the quadratic approximation on the same coordinate axes.
g) Use your approximation to estimate how long it would take for a golf ball to travel 150 feet.

time/sec	dist/feet
4.56	10
6.57	20
8.14	30
9.48	40
10.65	50
11.85	60
12.46	70

Graphing Help.

The following are some different ways to generate the graph of the points and curve using the same coordinate axes.

Maple:

```
> with(plots);
> points := [[0, 2], [1, 3], [2, 6], [3, 10]];
> plot1 := plot([points], style = point, symbol = solidcircle, color = black);
> plot2 := plot(2.7*x+1.2, x = -1 .. 4, color = black);
> display(plot1, plot2, view = [-1 .. 4, 1 .. 11]);
```

A couple observations: I added in the `symbol = solidcircle` command to make the points more visible. There are other symbols you could use. Also, the default color for these plots is red, so a change to black is nice for printing. I also used `view = [-1 .. 4, 1 .. 11]` instead of `view = [0 .. 3, 2 .. 10]` so that the you can see x and y values both smaller and larger than those in the data set. This allows the graph to be easier to read. Some versions of Maple will allow you to right click on the graph and set other options. Try.

Once you have generated the graph, I suggest you copy it and paste it into a word document. Students have had some trouble printing from maple in the past. Also, once in Word, you can resize it if you need to.

TI Calculators:

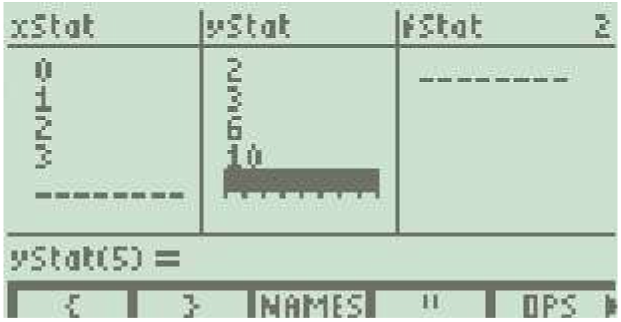
It is nice to be able to graph a set of points and a function with your TI. However, it is not easy to get a printout of your graph. One way would be to take a digital picture of the screen. Another way is to use the Virtual TI calculator program and a screen print utility like ScreenPrint32 to capture the TI screen image. You do need to have a ROM file from a TI calculator to use this virtual TI program though.

In any case, here is how you can graph our first example with a TI 86:

From a blank screen (hit exit and clear a few times) enter the list screen, 2nd Minus

. Then F4

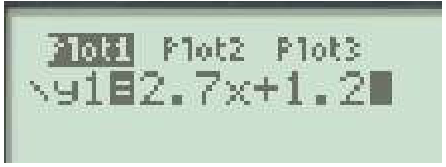
for the EDIT menu. Hit 0 and then hit Enter. This is the first x value of our data list. Then 1 enter, 2 Enter and 3 Enter. Now arrow over to the right once. Hit 2 and enter. This is the first y value from our data. Then 3 and Enter, 6 and Enter, and 10 and Enter. Your screen should look like this:



xStat	yStat	fStat 2
0	2	--------
1	3	
2	6	
3	10	
--------		

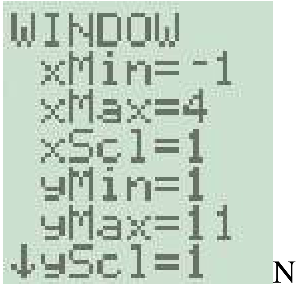
yStat(5) =

Now hit Exit so you are to a clear screen. Next we will enter our function: Hit Graph, then F1 for y(x)=. To get our data to plot, arrow up once and hit Enter. The Plot 1 option should now be dark. Arrow back down and enter 2.7x+1.2 Your screen should look like this:



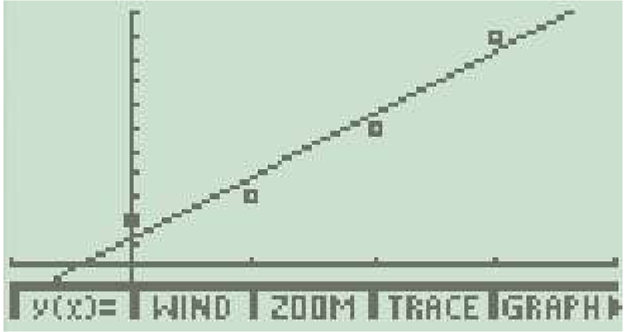
```
Plot1 Plot2 Plot3
\y1=2.7x+1.2
```

Next, let's set our window. Press 2nd F2 and we see our window screen. Let's use similar x and y values as in the Maple example above, so hit -1 Enter, 4 Enter, Enter, -2 Enter, 11 and Enter. Here is the screen:



```
WINDOW
 xMin=-1
 xMax=4
 xScl=1
 yMin=1
 yMax=11
↓yScl=1
```

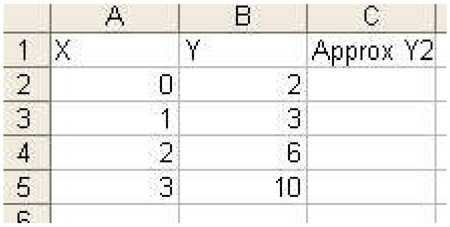
Now hit F5 to graph. We had to drop down to YMin = -2 so we could see the final graph, and here it is:



Other models of TI calculators have a similar setup. If you need extra help, use your manual, office hours, other students, google, etc. All of the above pictures were captured using ScreenPrint32 and Virtual TI with the TI86 rom and then inserted into this word document.

As a last step, you should go back into the y(x)= screen, and unselect the Plot1 option. Otherwise you will have problems with regular graphing.

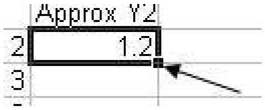
Excel:



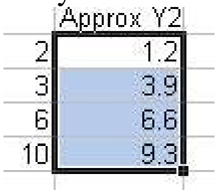
	A	B	C
1	X	Y	Approx Y2
2	0	2	
3	1	3	
4	2	6	
5	3	10	

Starting from a blank worksheet, enter the following: . These are the initial data, including labels in the first row. Now, in the cell C2, enter the following: =2.7*A2+1.2 Then hit enter. The cell should change to 1.2. This tells this cell to take 2.7 times the x value, and add 1.2, that is 2.7x+1.2, our approximate model. On the problems with squares, use A2^2 for x^2, etc.

Now select this cell by clicking on it and you will see a small black square on the bottom right corner: Click and hold on this square and drag down until the shadow outline covers three more cells. If you make a mistake, you can always hit ctrl+z to undo the step, and try again. You should now have this:

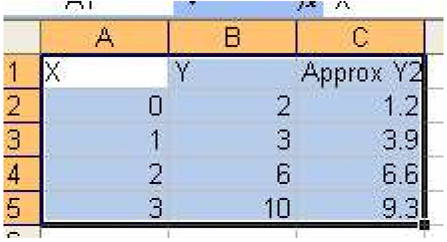


	Approx Y2
2	1.2
3	

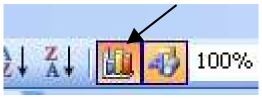


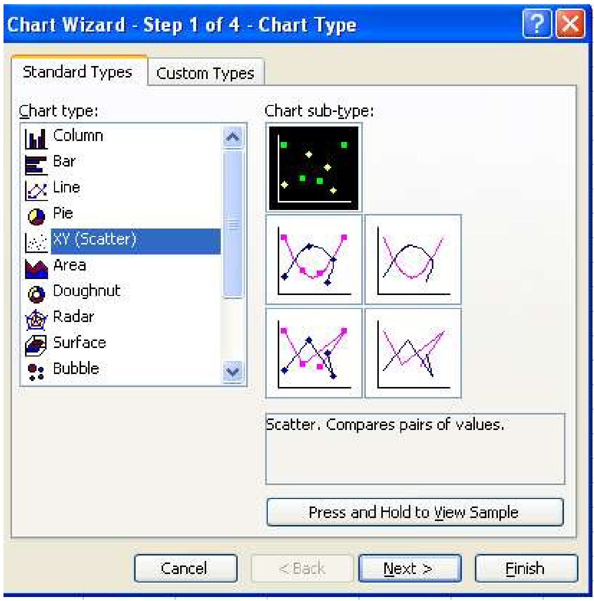
	Approx Y2
2	1.2
3	3.9
6	6.6
10	9.3

You have copied the approximation formula to these cells. Next, select from A1 to C5 so all of your data and labels are in blue: and click on the chart wizard button: and choose the scatter plot option and then finish:



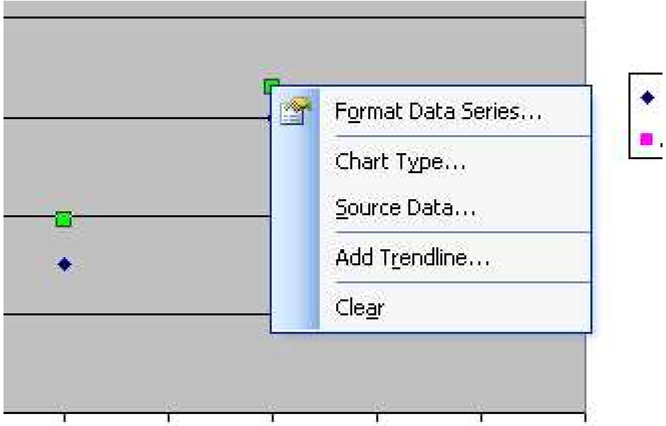
	A	B	C
1	X	Y	Approx Y2
2	0	2	1.2
3	1	3	3.9
4	2	6	6.6
5	3	10	9.3



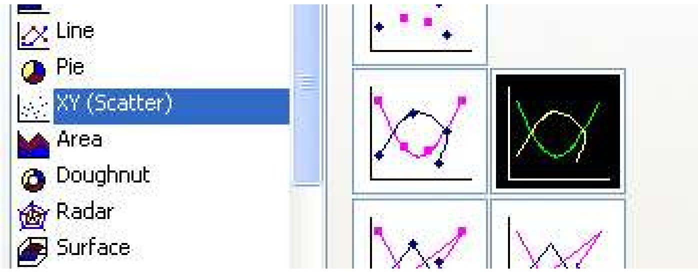


Scatter. Compares pairs of values.

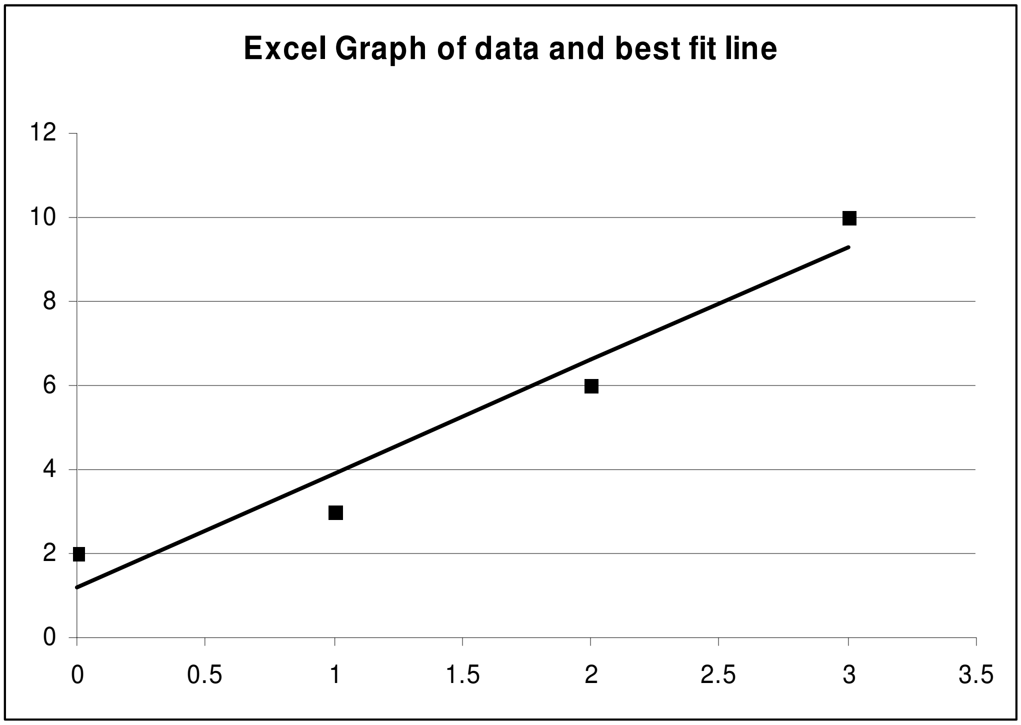
Now, right click on one of the purple squares (they will turn green) and choose chart type:



Then choose the right middle option of the XY(Scatter):



and click OK. You now have the graph. By right clicking on different parts of the graph, you can change colors of the line or points, delete the legend, edit the grid lines, add a title, etc until you are satisfied. Then select the whole chart (click near a corner), copy and paste into word to get, for example:


Excel Graph of data and best fit line