10. LECTURE 10

Objectives

- ☐ I understand the difficulty in finding an appropriate function for a data set in general.
- ☐ In some cases, I can define a function type that may fit a data set well.

Last time, we learned the procedures for how to fit a particular function to a data set. Where does such a function come from? If we have a data set, how can we determine which function(s) to try?

Unfortunately, there is no good answer to this question. Throughout the sciences, researchers have defined important relationships between physical measurements and come up with mathematical expressions like those below.

- **Law of Gravitation:** $F = G\frac{m_1 m_2}{r^2}$
- **Boyle's Law (pressure & volume):** $P_1V_1 = P_2V_2$
- **Hardy-Weinberg Law (population genetics):** $(p+q)^2 = 1$

Generally speaking, these laws are discovered is by lots of observations and by trial and error. The functions tried are often based on researchers' expectations. Once a consistently good equation is found, researchers then attempt to justify why such an equation makes sense for the situation.

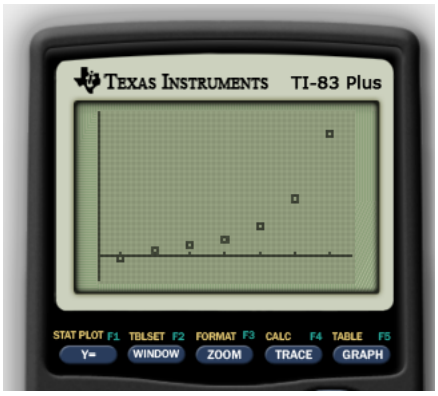
Here, we will spend some time discussing how we might determine a function for a data set with no physical intuition.

10.1. **Combining Different Fits for One Independent Variable.** Suppose you have the following data set:

x	y
1	-11.32
2	12.85
3	27.69
4	46.65
5	81.10
6	157. 19
7	344.44

I have generated this data from a particular function. Let's see if we can guess it.

The first step is to graph the data and study the general shape. Below is a picture of the data.



Given its shape, we might begin by trying to fit an exponential equation. If we try to calculate a best-fit with `ExpReg`, we get an error message.

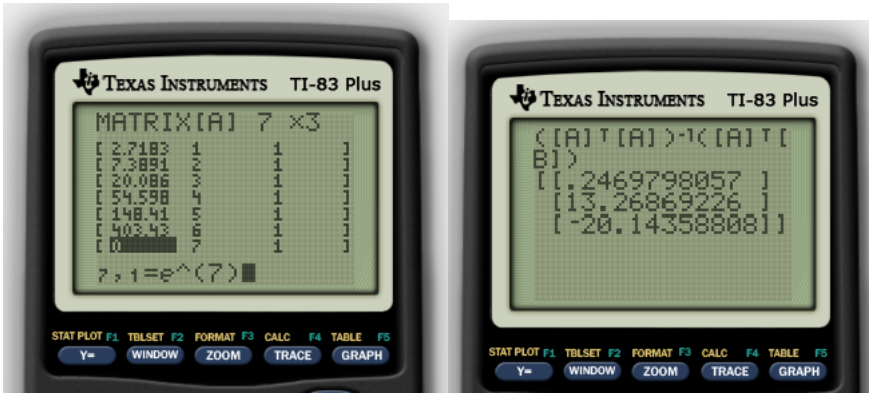
The error is an issue with the domain. That means one of the y-values is not appropriate for the type of function. This would be -11.32.

We may suppose that it is an exponential shifted down by a constant (about -8.6). A quick calculation shows us that if this were the case, then the y-value for $x = 2$ would also be negative. Therefore, we must have an additional term. We are not sure what term this might be, so we may try a linear one.

That gives us the form

$$y = ae^x + bx + c.$$

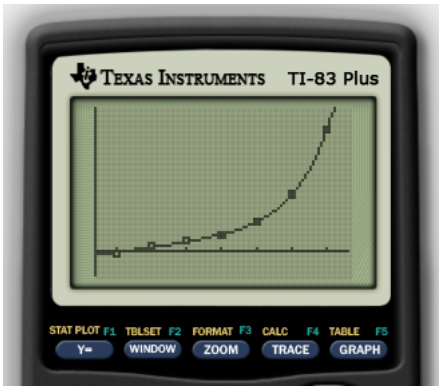
If we fit this function to the data, we get the following:



That gives us the function

$$y = .25e^x + 13x - 20.$$

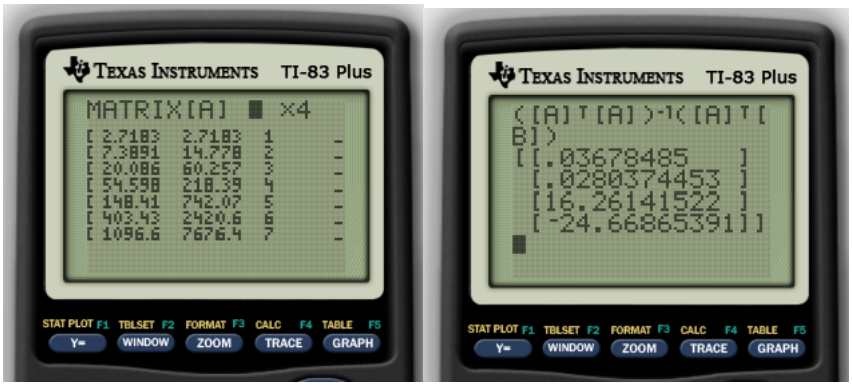
Let's see how well this matches the function.



This is actually quite good, but we may be able to do better. Let's consider a cross-terms between x and e^x. That is,

$$y = axe^x + be^x + cx + d.$$

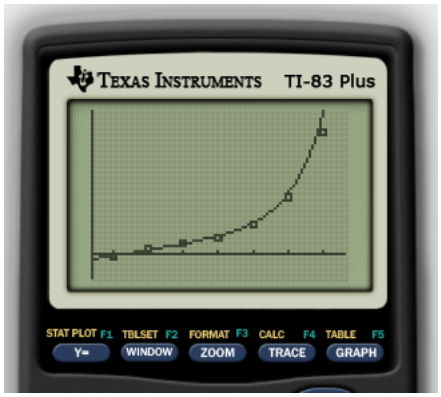
If we fit this function to the data, we get the following:



That gives us the function

$$y = .037xe^x + .028e^x + 16x - 24.$$

Let's see how well this matches the function.



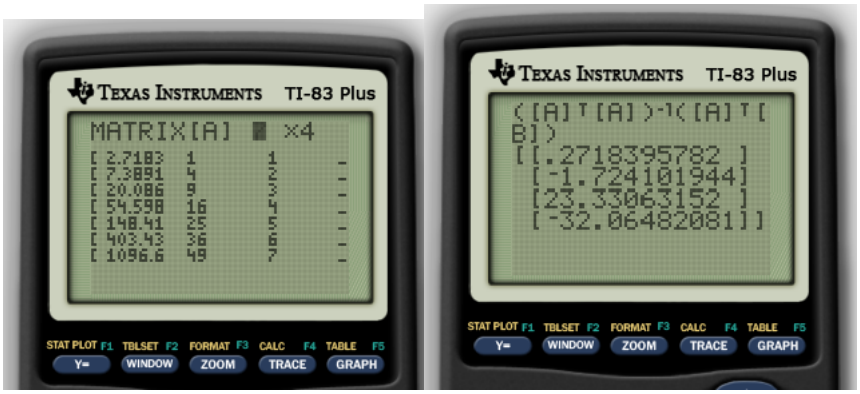
This fit is worse than the first. Notice that the curve is increasing faster than the points. That suggests that xe^x might be a bad term to add in since it is the fastest growing term.[7]

[7] You can take the derivative to verify this.

Another possible function to consider is one with a higher-order polynomial term. That is,

$$y = ae^x + bx^2 + cx + d.$$

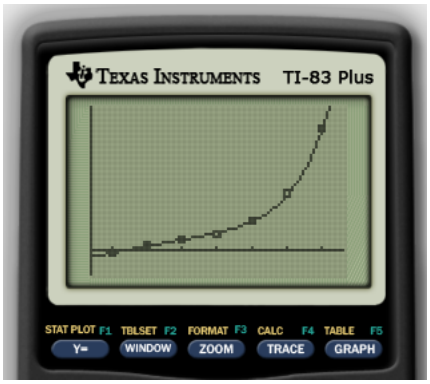
If we fit this function to the data, we get the following:



That gives us the function

$$y = .27e^x - 1.7x^2 + 23x - 32.$$

Let's see how well this matches the data.



This is pretty good. We can, of course, keep going and test a variety of other functions next. For example,

- $y = ae^x + bx^3 + cx^2 + dx + e$
- $y = ae^x + bx^2 + cx + d\sqrt{x} + e$
- $y = ae^x + bx^2 + cx + \frac{d}{x} + e$
- $y = ae^x + b\ln(x) + cx^2 + dx + e$

Any of these would be worth trying, but it is not very clear when one has the best possible answer. What is a great fit?

For this data set, it turns out the numbers come from the equation

$$y = .25e^x + x^2 + 27 - \frac{40}{x}.$$

If you study what we did, we were able to figure out the largest term e^x, but beyond that, we had a lot of trouble. This is pretty typical.

"Procedures" to Fit Data (One Variable) 10.1

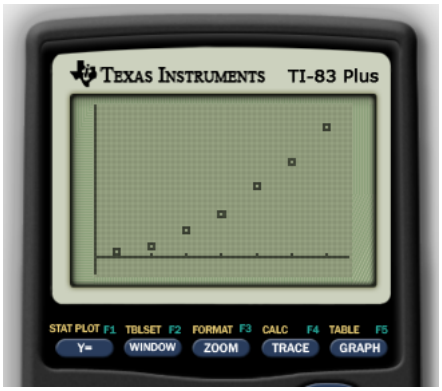
1.) Graph the data and try to determine a basic function type that may describe the general trend.
2.) Add in extra terms based on what you see and what you think may improve the fit.
3.) There is no clear stopping point. There is no clear way to verify that you have a good fit.

10.2. **Combining Different Fits for Many Independent Variables.** For multiple variables, it's a little more complicated. The idea is to consider each independent variable by itself. Then, once we have a set of functions, we will try to combine all of them into a multivariable function.

For example, suppose you have the data

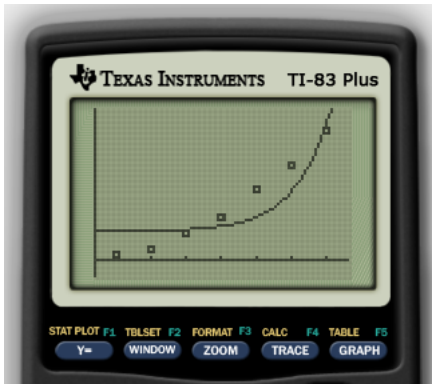
x_1	x_2	y
1	2	2.7
2	1	5.0
3	4	14.2
4	3	21.4
5	6	35.0
6	5	46.7
7	8	64.6

Then our first step is to find a function that models how x_1 influences y. As we saw above, that begins with a graph.



How might we describe this function? This looks exponential, so let's try fitting an exponential of the form $ae^{x_1} + b$.

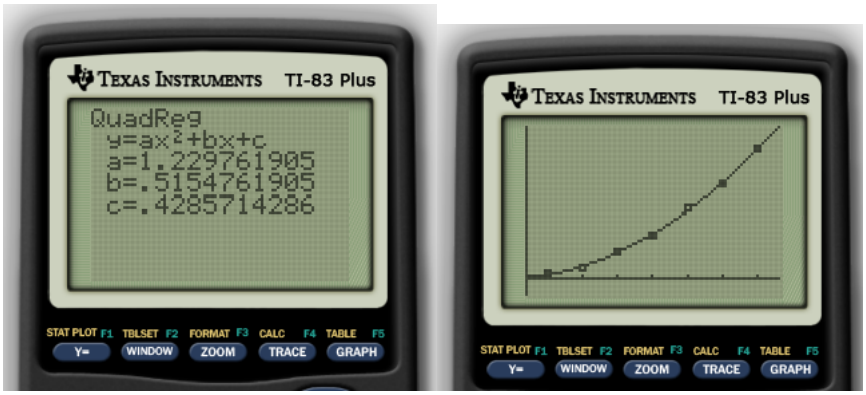
When we do, we get $y = .051e^{x_1} + 14.4$. If we compare this with the graph, we get the image below.



As you can see, it's a terrible fit! Our next attempt might be some other curved, increasing function like a quadratic. For that, we can use the calculator's program. It gives us the following function.

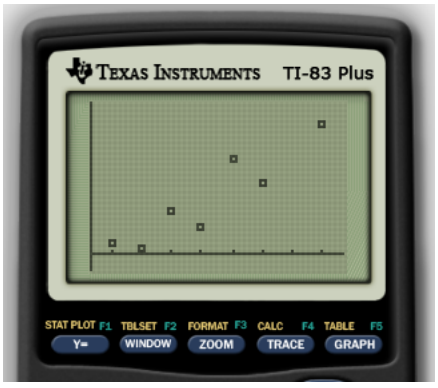
$$y = 1.3x_1^2 + .52x_1 + .42$$

Let's see how that holds up:

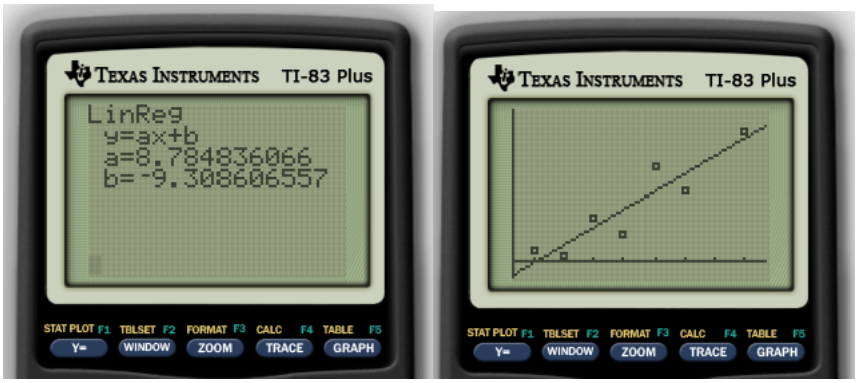


That looks quite good. So we will say that the relationship between y and x_1 is defined by a *quadratic* relationship.

Now let's see how y is related to x_2. First, we graph.



It's difficult to determine from the picture what kind of relationship exists. The only reasonably guess is linear. So let's find a best-fit line and compare.



Overall, this looks like a good fit. So we will assume that x_2 and y share a linear relationship. Therefore, our function will be constructed from these two function types.

That is, we will combine

$$y = ax_1^2 + bx_1 + c \quad \text{and} \quad y = ax_2 + b$$

by looking at all possible combinations of terms. We consider each term in one equation, like x_1^2, multiplied by each term in the second equation, $x_1^2x_2$, and by itself x_1^2. That gives us the function.

$$y = ax_1^2x_2 + bx_1x_2 + cx_1^2 + dx_1 + ex_2 + f$$

It is very possible that this function is way more than we need. So, we will try to fit it to the data. If any coefficient is zero or very close to zero, we will throw out that term.

In other words, we get the best-fit curve

$$y = -.006x_1^2x_2 + .007x_1x_2 + 1.2x_1^2 + -.06x_1 + 1.1x_2 + -.74$$

Now, we throw out the terms close to zero and try to re-fit the data. That is, we should instead consider

$$y = ax_1^2 + bx_1 + cx_2 + d.$$

When we do, we find the following coefficients.

In order words, our best-fit function for this data is

$$y = 1.1x_1^2 + .26x_1 + 1.1x_2 - .94$$

The actual function from which I got these numbers was the one below.

$$y = x_1^2 + x_1 \ln(x_2) + 1$$

It's not a perfect prediction. In reality, it is very hard to find the actual relationships. This is why, for example, it took humans a long time to realize that the sun was the center of the solar system. The geocentric models explained a lot of observations well enough. That is, until we developed a better telescopes and a better understanding of related topics, like gravity. Then, discrepancies led to finding a better model.

In the case of these two function, they don't differ too much if x_1 and x_2 are within the ranges of 1 through 7. But they will differ a lot for values far away from this range. Hence we are limited in how we use these functions for prediction. We cannot construct perfect predictions using just math. Having an understanding of a physical system is crucial to choosing better models.

In other words, expertise and mathematical know-how is *much* better than just mathematical know-how.

"Procedures" to Fit Data (Mulfi-Variable) 10.2

1.) Pick one independent variable.
2.) Graph the dependent variable against that independent variable and determine a basic function type that may describe the general trend.
3.) Add in extra terms based on what you see and what you think may improve the fit.
4.) Once you are satisfied with the function type you have selected, repeat this process for another independent variable.
5.) After you have defined one function-type for each independent variable, consider all possible combinations of the terms in all the functions. For example, if you have

$$y = a\ln(x_1) + bx_1 + c \quad \text{and} \quad y = ax_2 + \frac{b}{x_2} + c$$

then combine the two by considering each possible combination

$$y = ax_2\ln(x_1) + \frac{b\ln(x_1)}{x_2} + cx_1x_2 + \frac{dx_1}{x_2} + e$$

6.) Fit this multivariable function to the data.

10.3. **Next Time.** Up to this point, we have discussed how observations can give rise to single and multiple variable functions. From this point onward, we discuss how we use calculus to analyze these higher-dimensional functions.

Summary of Ideas: Lecture 10

- There is no standard way to find a good function for a data set. Ideally, you should use your knowledge of the physical system to guide you.
- You should also study the graph and try a number of possible functions.
- For multiple variables, you'll need to evaluate each independent variable separately and then combine your findings at the end. If any coefficient is close to zero, you should probably throw it out.

PENN STATE UNIVERSITY, UNIVERSITY PARK, PA 16802