Modeling Discrete Data: Introduction to the Issues

Most of us were introduced to mathematics through counting. A typical everyday problem that we learned to solve is: If a school bus holds 50 children, how many buses are required to take three classes, of 30, 35, and 32 students each, on a school trip? Such problems taught us how to add, subtract, multiply, and divide. Now, we can readily calculate that 2 buses will be needed.

Then came algebra. Algebra taught us to work with symbols: If Jeff is half as old as his father, and the sum of their ages is 60, how old is Jeff? We all know what to do. We could let x be the age of the father, and y be the age of Jeff. Then the information leads to two equations that yield $x = 40$ and $y = 20$.

Now, we are going to begin the study of calculus. Calculus gives us the tools to answer questions about movement and change, while building on our knowledge of arithmetic and algebra. A typical example of the kind of question we will learn to answer is the following: Suppose we drop an object vertically, from rest, off the roof of a building. How long will it take for it to hit the sidewalk ten meters below?

Just as we have been trained in the arithmetic and algebra problems above, the first thing we must do is translate the problem into mathematical language. Because we are just finding our way, this is not so easy. But it appears that as the question now stands, we need some additional information. For example, it would be nice if we had a formula that would give us the time corresponding to any distance of fall. Even though this may seem like too much to ask for, we will see that this indeed will be one of the outcomes of our study of calculus. But to jump to that formula now would be like pulling a rabbit out of a hat. We want to understand how we get it, and where it comes from, so that we can apply the same or similar principles in other situations. Thus, we will take smaller steps.

Suppose for example that a similar experiment has been conducted in a lab where the distances (in meters) have been recorded every tenth of a second for one second, as in the table below. If we could extrapolate the data to 10 meters, then we would be able to answer the question about the object falling from the roof. So, how shall we proceed?

time (s)	distance (m)
0.10	0.049
0.20	0.196
0.30	0.441
0.40	0.784
0.50	1.225
0.60	1.764
0.70	2.401
0.80	3.136
0.90	3.969
1.00	4.900

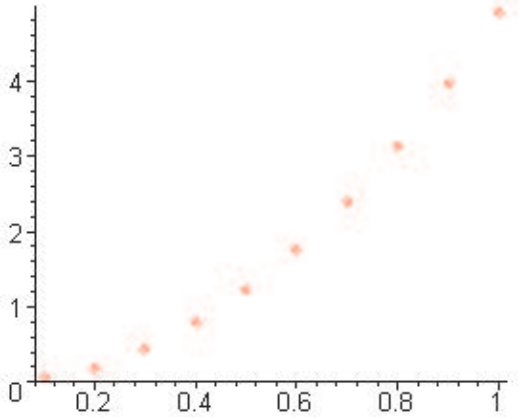
A reasonable approach is to try to find a function that will model the data. That is, find a function $f(t)$ of time t whose values at the recorded times come close to matching the distances in the table. Then we can use that function to find the desired time, namely, by solving for t in the equation $f(t) = 10$. But how do we find such a function, and how do we know which is best if we find more than one?

Let's postpone the question of deciding which function is best among several choices. So far we don't even have one candidate; hence, let's deal with that problem first. Building on what we know, a good place to start our search is with the *elementary functions*. After all, these functions – polynomials, exponentials, logarithms, trigonometric functions – in addition to being familiar to us, have proved themselves throughout the years to be very valuable in just such situations.

0.1 Modeling with an Elementary Function

We will start with the least complicated elementary function and continue testing until, hopefully, we find one that fits the data fairly well. Eventually we will have to become more precise about concepts such as *fairly well*, but for now we will forge ahead just to see if the plan holds any promise at all.

The data certainly are not constant. Thus, we check to see if they are linear. That is, is there a function of the form $f(t) = at + b$ that fits the data in the table? Before going further and substituting points in an effort to find a and b, perhaps we should plot the data to see what their graph looks like. Here is a plot:



The plot clearly shows that the data do not fall on a (straight) line. We could verify this numerically from the table by showing that the slope changes as we move from one point to the next. Instead, in our search for an elementary function, we will trust our eyes and move on to quadratic functions; that is, functions of the form $f(t) = at^2$. (As a first step, we have decided not to use the most general form of a quadratic. If this does not yield good results, we will next try a more generate quadratic function whose graph is not symmetric about the origin.)

Substituting $t = .1$ and $at^2 = .049$ gives $a = 4.9$. Thus, the candidate is now $f(t) = 4.9t^2$ and we check it against other values in the table: $f(.2) = 4.9 \times .2^2 = 0.196$, $f(.3) = .441$, etc. Amazing! This function fits the data of the table exactly. (You might want to substitute the other values of the table to assure yourself of this fact.)

Now, we can complete our plan and answer the original question: The time it takes the object to fall from rest a distance of ten meters is found by solving the equation $4.9t^2 = 10$. Thus, the desired time is about 1.43 seconds.

0.2 Review of Modeling Issues So Far

This is a good place to review the key points of our work so far, before moving on to more complicated situations.

We began with a typical question about motion that calculus was developed to address. Rather than go right to the solution, we have approached the problem from the standpoint of someone who does not know calculus. Our aim is to identify those elements of the problem that will be crucial in our future development of calculus and its tools. The first key thing to note is that real-world problems very often involve discrete data. This is true even though the underlying process may be continuous. When we translate the problem into mathematical terms, our first attempt usually involves some finite number of discrete measurements.

In the example we have considered of the object falling off a roof, we had the benefit of laboratory data. Someone had to go to the lab and conduct a more controlled experiment where it is possible to measure the distances that correspond to specific instants of elapsed time. We don't know the form of the function that lies behind the data. The best we can do is to observe and record positions at a finite number of times.

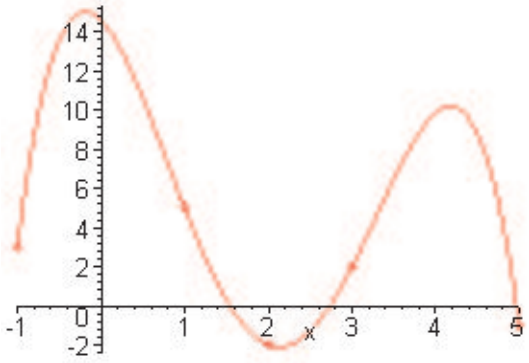
Applet: Falling Object **Try it!**

The next key thing to note is that there is usually an underlying function that fits the data and also

takes on values at intermediary points. Furthermore, graphing the data is an important aid in identifying the function. So, we should expect functions and graphs to play major roles in our work modeling problems of motion. In fact, our mathematical modeling will take the form of asking the following question: Do we know of a mathematical function that fits the data and fills in intermediary points in a way that is consistent with the system we are observing?

Invariably, the answer to that question will be an elementary function, or a function related to an elementary function. We will be seeking a function by trial and error that comes as close as we can to hitting the data points.

You probably already know that given a finite number n of points, it is always possible to find a polynomial of degree at most $n-1$ that passes through them. However, this function usually is not satisfactory for modeling purposes because of its behavior either between the data points or beyond them. For example suppose that we want a polynomial that passes through the points (-1, 3), (1, 5), (2, -2), (3, 1), and (5, -1). Since the general 4th degree polynomial $p(x) = ax^4 + bx^3 + cx^2 + dx + e$ has 5 unknown coefficients, we can expect to choose them in such a way that the 5 equations $p(-1) = 3$, $p(1) = 5$, $p(2) = -2$, $p(3) = 1$, and $p(5) = -1$ are satisfied. Solving these equations does indeed produce solutions for a, b, c, d, and e, and we graph the resulting polynomial $p(x)$ below.



In scientific applications it is rarely interesting to find a polynomial that passes *exactly* through a given set of data points. This is because the data are themselves usually approximations, and so finding a polynomial of degree n that passes through the n points is ascribing to the degree of the polynomial a meaning that has no scientific basis.

Applet: Falling Object **Try it!**

We will return to this issue in an example below involving census data. When we examined the lab data for the falling object, it was so precise that we were able to find a quadratic polynomial that fit the 10 points exactly. But suppose the points were merely close to the values of the quadratic and not exactly the same. We probably would still choose the quadratic over the 9th degree polynomial that passes exactly through the points. This is true because we have the sense that the quadratic works just as well for the intermediate points that correspond to the underlying continuous process. Hence, we want to emphasize this consideration when we choose a fitting function: It should work just as well at intermediate points, and at reasonably near points beyond the range of the data. The meanings of *just as well* and *reasonably near* will depend on the problem and the nature of the data we are attempting to model.

Moreover, there may be some characteristics of the physical system that inherently favor one function over another. We will see later when we develop some calculus ideas that this is the case for the falling object. We will be able to confirm that our choice of the quadratic function is correct on theoretical grounds.

0.3 Fitting a Curve to Data: The Method of Least Squares

Most data that comes from real-world situations is not fit exactly by any function of interest. This is primarily the case because the data themselves are only measured estimates. Take, for example, the US population census data (in thousands) for the years 1790 - 1850.

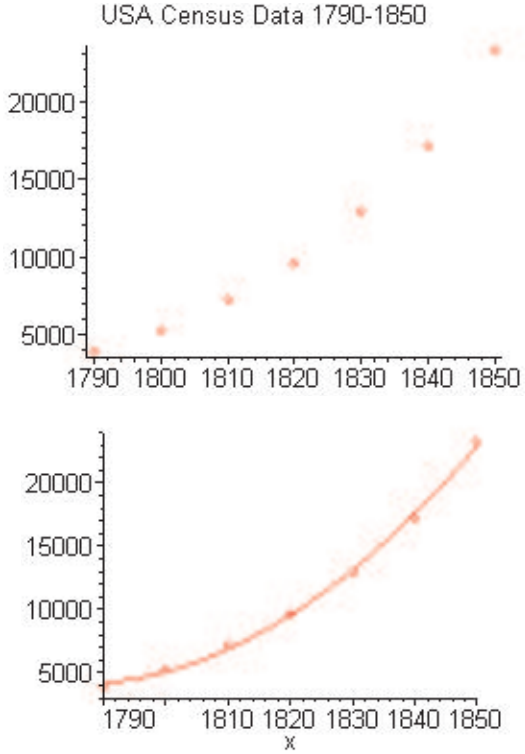
USA Census Data (thousands)	
Year	Population
1790	3929
1800	5297
1810	7224
1820	9618
1830	12901
1840	17120
1850	23261

Let's fit a quadratic polynomial to the census data. But how do we do this? The most common approach, and the one implemented by computer algebra systems such as Maple, Mathematica, and Mathcad, is to use the method of Least Squares. In general, suppose we start with data points (x_1, c_1), (x_2, c_2), ..., (x_n, c_n) and we want to fit the data with the mth degree polynomial $y = p(x)$, $m \leq n-1$. Then we will define the Best Least Squares mth degree fit (m and n given, $m \leq n-1$) to be the polynomial of degree less than or equal to m that minimizes the sum of the individual squared errors, namely,

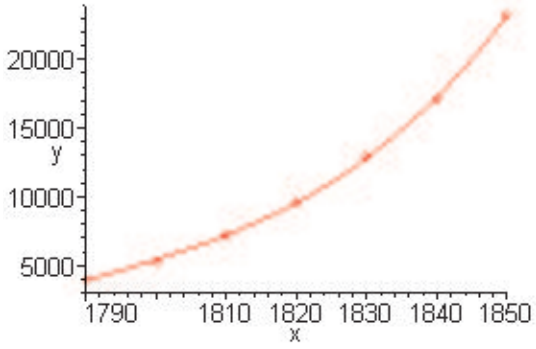
$$(p(x_1) - c_1)^2 + (p(x_2) - c_2)^2 + (p(x_3) - c_3)^2 + \ldots + (p(x_n) - c_n)^2 = \sum_{i=1}^{n} (p(x_i) - c_i)^2$$

where we have used so-called sigma-notation to write the summation in a compact form. That is, the outcome of a computer implementation of the Least Squares method will be to determine the coefficients of the polynomial $p(x)$ that will minimize the sum of the squared errors. Squaring the errors does not allow the signed-errors to cancel one another in the summation; sqaring also has the effect of giving more weight to bigger errors than smaller ones.

Here are the results of using Maple to find the Least Squares quadratic fit of the census data. The quadratic polynomial is $q(x) = 4.417023810x^2 - 15766.11310x + .1407294507 \times 10^8$, and below we show its graph with the original data points.

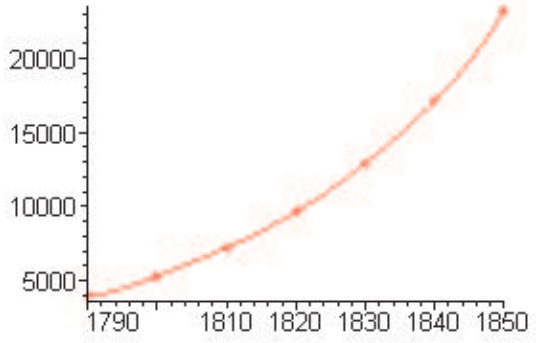


Visually, the fit looks pretty good. However, why stop with a quadratic? Next, let's fit a cubic polynomial to the data points. The Least Squares cubic fit is the polynomial $p(x) = .05097685185x^3 - 273.9165873x^2 + 490765.3753x - 293179528.2$; its graph also shows a good fit of the data.



Speaking again from a strictly visual perspective, the cubic fit actually looks *better* than the quadratic. In fact, we *know* that the cubic yields a smaller sum-of-squared-errors than the quadratic because in the Least Squares procedure, a quadratic function is always a possible outcome of fitting a 3rd degree polynomial to the data. Therefore, the sum of the squared errors for the cubic must be smaller than the sum of squared errors for the quadratic. In fact, given that we have formulas for the two polynomials, we can calculate directly the sum of squared errors for each. When we do this, we find them to be about 606852 for the quadratic and 47455 for the cubic. So, the cubic fit is considerably better according to this criterion.

But where does this leave us? First, we now have a mathematical criterion for deciding which of two different functions gives a better fit to a set of discrete data, namely, the one that has the smaller sum of squared errors. The essential consideration here is to be consistent with the criterion applied by the curve fitting computer algorithm we are using. Second, this criterion alone will not tell us which function to use to model the data. After all, with these 7 data points, we could always find the polynomial that pases through all of the points exactly and hence the sum of squared errors would be 0.



The 6th degree polynomial graphed above does just that: it hits every one of the 7 census points exactly. Moreover, it seems to give just as much of the proper feel at intermediate points as the cubic or the quadratic least squares fits. And yet we know that census data is inexact by its very nature; hence, it seems inappropriate to put the emphasis on obtaining an exact fit. Then, how do we decide what function to use to model the data? It is clear that we have to bring some other consideration to bear on the problem to decide among competing curves. For example, in the present case of the census data, it turns out that there are good theoretical reasons for the best fitting function not to be a polynomial at all, but rather an exponential. However, since we have not discussed exponential functions yet, we will have to postpone the completion of our analysis of the census data until we are familiar with exponential functions and their properties. We have reached a point in our discussion where we need to develop some new mathematical skills before we can continue further.

Applet: Least Squares Fitting **Try it!**

0.4 Our Agenda for This Chapter

We opened this section with an introduction to the kinds of questions that calculus can answer. We found that calculus gives us the tools to analyze motion and change. It does this by providing a system for studying and transforming functions. The functions are the tools whereby we represent moving and changing systems. The entire process starts with a set of discrete data and finding a function to model it.

Ultimately, calculus gives us the vocabulary to pose fundamental questions about a system, and the means to answer them. The elementary functions provide us with a robust library of functions which are sufficient for many of the purposes of calculus, enabling us to match the properties of the function with the observed behavior of the physical system. These *observables* are in the form of discrete data. Thus, the first step in modeling problems in calculus is to fit a (usually elementary, or related) function to the data.

The method of least squares is a powerful tool for determining the best fit. We normally do not seek a function that passes through all of the data points exactly because a set of data rarely describes an underlying system completely. The data are subject to experimental error and reflect other factors (such as friction in a moving system) that we do not choose to (or cannot) take into account. Thus, a function that passed through all of the data points would put too much emphasis on the data themselves and not the system. It is difficult to include all of the variables in a model because then the model can become too complicated to analyse. Thus, we look for both a mathematical reason and a systemic reason to choose our modeling function to be the simplest possible, making simplifying assumptions about our system along the way. The rest of the modeling involves studying the function and using the results to predict the behavior of the system.

We have raised many issues here, and have incidentally identified many gaps in our knowledge. In this section, we will start to fill them by:

1. Studying functions and their graphs.
2. Studying the elementary functions and their properties.
3. Getting more experience with modeling real discrete data.

All too often, the study of calculus appears to consist of a long list of rules (the derivative of this, the integral of that). In actuality, calculus begins with rates of change of functions that model discrete data and describe elementary geometric shapes (e.g., lines, parabolas, circles). These modeling ideas are an indispensible part of the foundation of the subject, and form the rich and textured fabric that ties calculus to the moving and changing world around us.

Applet: Calculator: Values of Elementary Functions **Try it!**
Exercises: Problems **Check what you have learned!**
Videos: Tutorial Solutions **See problems worked out!**