MATH 221 PROJECT

Determining functional form from data

1. Purpose of this project

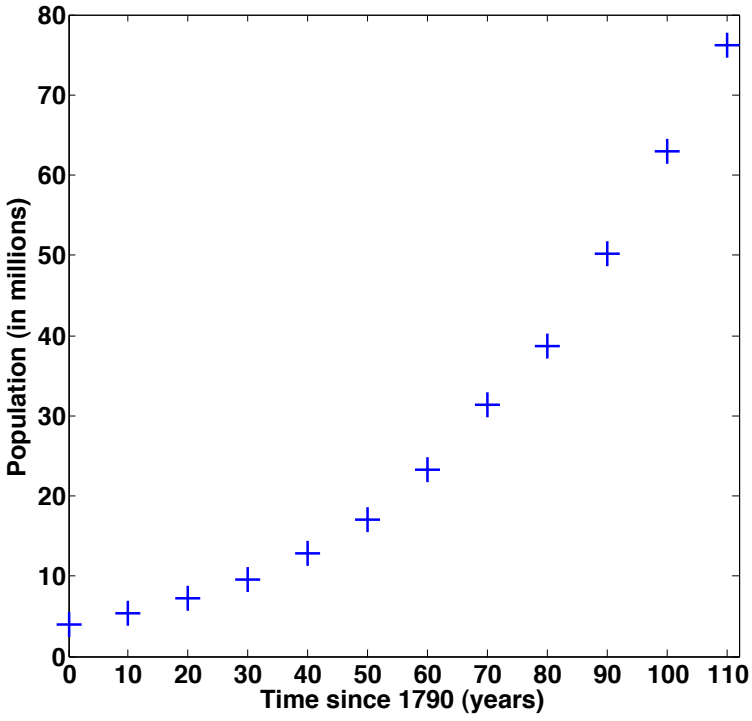
Suppose that you are presented with some data and want to figure out what type of function generated this data. For example, you may wish to better understand how the population of a certain region depends upon time, or you may wish to know how precise a particular method is for the numerical computation of derivatives.

In this project, we will learn to differentiate between data arising from a power function $f(t) = ct^k$ or an exponential function $g(t) = ce^{kt}$. Further, we will learn how to approximate the constants c and k from the data.

2. Case Study: U.S. Census Data

The table below provides population data (in millions) from 1790 - 1900 for the United States taken from http://www.census.gov/. It is not clear from the plot whether the data exhibits polynomial or exponential growth, or neither.

Year	Population (millions)
1790	3.9
1800	5.3
1810	7.2
1820	9.6
1830	12.9
1840	17.1
1850	23.2
1860	31.4
1870	38.6
1880	50.2
1890	63.0
1900	76.2

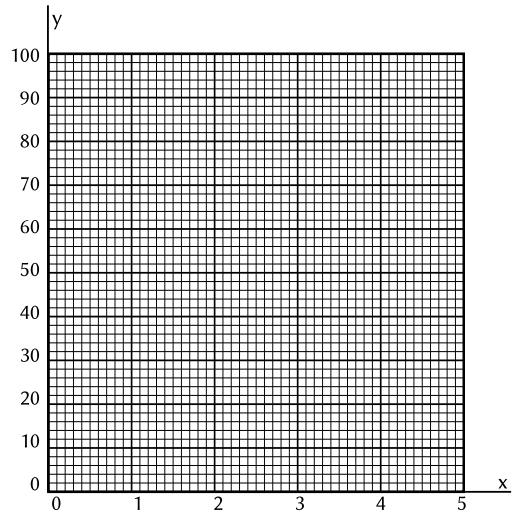


We begin by generating data from a known exponential function so that we can learn how to "discover" the original form of the function. Next, we will apply these techniques to real data.

3. Exponential functions

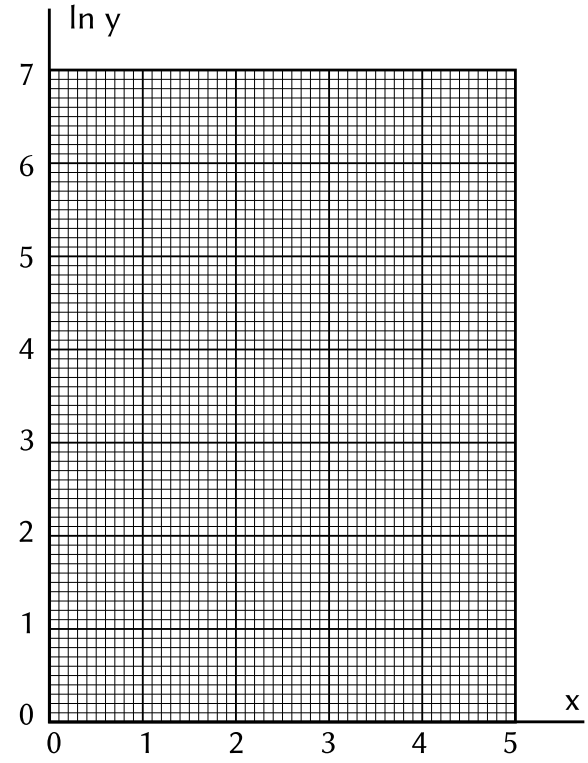
Consider the function $y = 3 \cdot 2^x$. We will use this function to generate some data. Fill in the table below for the indicated values of x. Next, plot the points (x, y) on the provided graph.

x	y
0	
1	
2	
3	
4	
5	



Now, using the values above and a calculator, fill in the table below and plot the points $(x, \ln y)$.

x	$\ln y$
0	
1	
2	
3	
4	
5	



4. What just happened?

You should notice that your second plot above appears to be a line. The second plot is usually called a "semilog plot" since it plots the natural logarithm of the y coordinate versus time. Should we have expected

the semilog plot to give a line? The answer is yes since taking logs of both sides of the equation $y = 3 \cdot 2^x$ yields

$$\ln y = \ln 3 + \ln(2)x.$$

Thus, treating $(\ln y)$ as a variable, we see that there is a linear relationship.

5. More generally

In general, suppose that

$$y = c \cdot e^{kx}. \tag{1}$$

For example, if $y = 3 \cdot 2^x$, then we can write $y = 3 \cdot e^{(\ln(2))x}$. Taking ln of both sides of equation (1) yields

$$\ln(y) = \ln(c) + kx.$$

Thus we may conclude the following.

Fact: ***if*** $y = c \cdot e^{kx}$ ***for some*** c ***and*** k***, then the plot of*** x ***versus*** $ln(y)$ ***will be a line.***

More can even be concluded! Not only does the linearity of the semilog plot tell us that the function satisfies an exponential form, but it also gives us a method for determining c and k since

(1) $\ln(c)$ is the vertical intercept of the semilog plot when $x = 0$, and
(2) k is the slope of the semilog plot.

6. Power functions

Now suppose that we are dealing with a polynomial function such as

$$y = c \cdot x^k,$$

for some real numbers c and k. Will a similar idea work? Taking ln of both sides of the above equation yields

$$\ln(y) = \ln(c) + k \ln(x).$$

Therefore, we see that plotting $\ln(x)$ versus $\ln(y)$ will produce a line. Further, both c and k can be recovered by finding the slope and intercept of the resulting plot. Such a plot is referred to as a log-log plot since now both the dependent and independent variables have logarithms applied to them.

7. Problem

1. Consider the function

$$f(x) = e^{-x^2/2},$$

which plays an important role in probability. We will numerically approximate $f'(0.5)$ using three different methods:

(1) The forward difference: $f'(0.5) \approx \dfrac{f(0.5+h) - f(0.5)}{h}$.
(2) The backward difference: $f'(0.5) \approx \dfrac{f(0.5) - f(0.5-h)}{h}$.
(3) The centered difference: $f'(0.5) \approx \dfrac{f(0.5+h) - f(0.5-h)}{2h}$.

Fill in the following tables with the relevant data:

h	$\frac{f(0.5+h)-f(0.5)}{h}$	$\frac{f(0.5)-f(0.5-h)}{h}$	$\frac{f(0.5+h)-f(0.5-h)}{2h}$
1/5	-0.498 961		
1/10			
1/50			
1/100			
1/200			

The exact value of the derivative is $-0.44124845\cdots$. In the table below, give the error induced by the different methods

$$E(h) = \left|f'(0.5) - \text{approximate derivative with parameter } h\right|.$$

Note the absolute values above!

h	$E(h)$	$E(h)$	$E(h)$
1/5	0.057 713	0.073 746	0.008 016
1/10			
1/50			
1/100			
1/200			

Produce log-log plots of the data in the three tables above and derive the functional dependence of the error for each method as a function of h. Can you conclude that one method is definitively "better" then the others? In what precise sense is it better?

8. Report Instructions

Your project writeup should include:

(1) Summary of ideas detailed in project above including a derivation showing why the semilog or log-log plots will give a linear relationship under different circumstances.
(2) Complete solution to Exercise 7.1.
(3) Complete the following two exercises:
   (a) Taking the data from the US Census on the first page of this project, determine if the population data exhibits exponential or polynomial growth. Drawing a line by hand through the resulting linear plot (this will be either a log-log or semilog plot), determine the relevant constants c and k. That is, find a time dependent function of the population of the United States that fits the data well. What does the model predict the population of the United States would be in the year 2010? Explain.
   (b) A numerical method used to calculate integrals produced the following errors, $E(h)$, where $E(h)$ is a function of a certain parameter h:

h	$1/10$	$1/20$	$1/25$	$1/40$	$1/50$
$E(h)$	0.071	0.0174	0.0114	0.00438	0.0027

   A scientist believes that $E(h)$ is well approximated by a polynomial of the form Ch^k, for some $C, k > 0$, where k is a positive integer. Is the scientist right? Explain. If the scientist is right, what are the values of k and C?