

A Model for World Population

In this lab, you will examine data about the population of the world and develop a mathematical model for human population from antiquity into the future. The technique used is called curve fitting. We will compare actual data--in this case, world population as a function of time--with mathematical models for how population changes with time. This kind of modeling is very common in the quantitative sciences, and this computer lab will help you understand the basic approach, its strengths, and its limitations.

Overview of lab:

A Prelab exercises

B Linear least-squares fitting with KaledaGraph

C Exponential fitting with KaledaGraph

D How good is a fit? Correlation coefficients

E Post hoc, ergo proctor hoc

F Modeling World Population as Exponential Growth

G A more sophisticated model for world population

G.1 Specific growth rate and human population

G.2 Finding specific growth rates during different historical periods

G.3 Modeling the World population from antiquity into the future

A Prelab exercises:

The most common type of numerical data is two dimensional, that is, it involves two experimental variables and hence can be expressed as x,y pairs. One experimental variable, by convention x , is called the independant variable, and it is usually the variable whose value the researcher can choose. In our population examples, the independant variable is time. The other variable, y , called the dependant variable, is what the researcher measures (e.g., number of people). In order to do curve fitting, we need a model that makes predictions about how the value of y changes for different values of x . These models take the form of mathematical functions which can be graphed in two dimensions

1) Your first task is to make thumbnail sketches of how the following functions look on the interval $x=0$ to 5. Hint: calculate the value of each function for several values of x .

Linear:
 $y=mx+b$, $m=2$, $b=-1$

Exponential:
 $y=y_0e^{(x-x_0)} + b$, $y_0=0.8$, $x_0=1$, $b=0$

Note that these functions all have a functional form with some parameters-- m , b , x_0 . For any given curve, the parameters have a constant value; you need to know this value in order to draw the curve. Changing the parameters changes the curve.

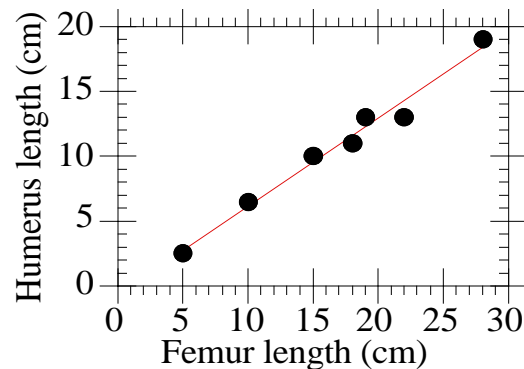
2) In the case of the linear example, to what features of the curve do the parameters 'm' and 'b' refer? Which parameter could you change, and how, to make the curve pass through the origin?

3) For the exponential curve, what happens to the value of y at very large and very small values of x ? Can you change the shape of the curve by changing y_0 ?

If a set of experimental data points are relatively close to the predictions of a mathematical model, we say the model "fits" the data, and that the data are related by the type of equation in the model. For example, if one measures the length of femurs (thigh bones) and humeruses (upper arm bones), one might get the following data:

Femur length	Humerus length
18.000	11.000
19.000	13.000
28.000	19.000
10.000	6.500
22.000	13.000
15.000	10.000
5.0000	2.500

Plotting femur vs humerus length (black dots) shows that the longer the femur, the longer the humerus.

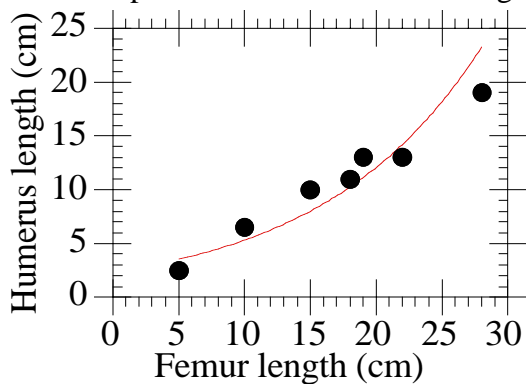


The line in the graph was generated by a computer program called Kaleidagraph. I told the program to fit this data as a line, with equation $y=mx + b$. The computer tried many values of the parameters 'm' and 'b' until it generated a line that was as close to all of the data points as

possible. This process is called “curve fitting.” Just by your eye, you can see that most of the data lies very close to the theoretical line (later we will learn more quantitative measures of how good a fit is).

Why is curve fitting important? First, the numerical parameters can be used to make predictions. An anthropologist can predict how tall a Neanderthal woman was from a single leg bone. More importantly, the fact that there is a linear relationship between these two bone sizes suggests some underlying rules about how bodies grow. One can then think of many measurements to test these rules, so curve fitting can give us ideas for new experiments.

Now, it is all well and good that the computer can juggle parameters and find the best curve, but one must always consider the GIGO principle (garbage in, garbage out). For example, what if I tell the computer to fit the same data using an exponential function?



The computer is just as happy finding an exponential fit to the data, even though there is no rational reason why such a relationship should exist. As you can see, the exponential model is a worse fit to the data than the linear one. Fewer data points lie on the line, and the 28 cm femur is way off the line. That’s because of a flaw in the model- it does not represent the underlying physical reality. ALWAYS CONSIDER THE MODEL!!!

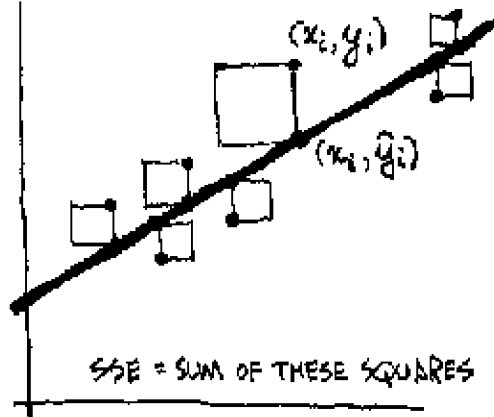
4) Why is the most obviously wrong data point the 28 cm femur? Hint--think about the difference between the two mathematical models for large values of x.

B Linear least-squares fitting with Kaleidagraph

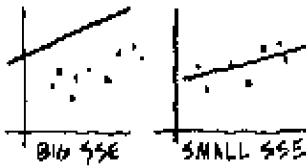
1. Sit down at a mac, and get a copy of Kaleidagraph. First, login to the computer, then select “Chooser” from the apple menu. In the top left box of the chooser, click on “AppleShare”, and then in the bottom left box, scroll down until you see “OIT Services” then make sure that “OIT Services” is highlighted. Once you have done this, the right box should say “Select a file server” above it. Double click on “ND Mac Server”, and when the dialog box pops up, log in as a “guest”. At the next dialog box, double-click on “ND Mac Applications”. At this point, an icon should appear on the desktop, labeled “ND Mac Applications”. Open this drive and scroll down until you see “KaliedaGraph 3.0”, copy this folder to the local hard drive by dragging it onto the “Macintosh HD” icon. Now, when you open “Macintosh HD” you should see the KaliedaGraph folder and you can start getting to work.
2. The next step is to find the data you will need to work on. Open the Courseware server (add more detail when we know exactly where the files will be), find the folder for Chem191 and open the file detailing the growth of the town of Podunk, listed as “**Podunk data**”. This is a text file and will be opened up in SimpleText. Also, open “Kaleidagraph”.
3. Next, enter the data that is in the text file into Kaleidagraph. Use two columns to enter the data, one for the number of houses and one for the year. To enter data, just click in a cell and type. To title a column, double click in the cell immediately above row 0. (just below the title bar) At the dialog box that comes up enter the column title and click “Done”.
4. Plot the data. Go to the “Gallery” menu, select “Linear” and then select “Scatter” from the sub-menu. At the dialog box that appears, click the radio button under “X” for the data you want to plot on the x-axis (the year, in this case) and click the radio button under “Y” for the data you want to plot on the y-axis. (Number of houses in this case) Then click “Done”.
5. Next, try to fit a line to the data. Go to the “Curve Fit” menu, select “Linear...” and at the dialog box, click next to the data set you want to fit a curve to. (In this case, number of houses) When you fit a line (or any other kind of curve) to a set of data, the computer does a calculation that finds the line that minimizes the sum of the vertical distances between all of the points and the line itself. These distances are then squared to eliminate problems with signs. So, the problem of curve-fitting can be visualized as an attempt to draw the line that will generate the smallest total area of squares, as shown in the diagram.

THE IDEA IS TO MINIMIZE THE TOTAL SPREAD OF THE y VALUES FROM THE LINE. JUST AS WHEN WE DEFINED THE VARIANCE, WE LOOK AT ALL THE SQUARED y DISTANCES FROM THE LINE, AND ADD THEM UP TO GET THE SUM OF SQUARED ERRORS:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



IT'S AN AGGREGATE MEASURE OF HOW MUCH THE LINE'S "PREDICTED y_i ," OR \hat{y}_i , DIFFER FROM THE ACTUAL DATA VALUES y_i .



The regression or least squares line

IS THE LINE WITH THE SMALLEST SSE.

The equation for the line generated by least squares analysis is: $y=mx+b$, where:

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad m = \bar{y} - b \bar{x}, \quad \text{where } \bar{y} \text{ and } \bar{x} \text{ are the means of the sets } Y_i \text{ and } X_i$$

6. Although a single line does not fit this data particularly well, two lines do fit the data rather well. Return to the data and find the point where the data should be broken up so that two straight lines can be used to fit the data. Once you have found this point, select all of the data in the number of houses column from that point downwards. Then, go to the "Edit" menu and select "Cut". The selected data should disappear. Fear not! Click in the third column (which should be empty) in the row across from the point where you broke the data up. (i.e. if you decided to cut the data after the fourth row, click in the fifth row of the third column) Then, go back to the "Edit" menu and select "Paste". Now, your data should reappear one column to the right.

7. Replot your data, this time putting "Year" on the x-axis and both "number of houses" and "C" (where your pasted data should appear) on the y-axis.

8. Try to fit two lines to the data. Go to the “Curve Fit” menu and select “Linear...”. At the dialog box that appears, check the boxes next to both “number of houses” and “C”, the computer will then fit separate lines to both sets of data. At this point, two lines should appear, one through each of the data subsets, and as you should be able to see this data set can be fit much more accurately by two lines than by one.

- What might have happened in or around Podunk to cause this change in growth of the town?

C. Exponential fitting with Kaleidagraph

1. Next, go back to the courseware server, open the text file, “**microsoft profits**”, and open a new document in KaleidaGraph by going to the “File” menu and selecting “New”.

2. Enter the data from the text file into the new file, like you did above, and plot the data with “year” on the x-axis and “dollars” on the y-axis.

3. As you can probably see this data looks like an exponential function. So, go to the “Curve Fit” menu and select “Exponential...”. The computer may or may not (depending on its mood) draw an exponential curve that fits the data. But, even if the computer decides not to draw the curve, it still will find the equation of the curve. To see this equation, go back to the “Curve Fit” menu, select “Exponential...” and at the dialog box click in the box below “View”. This will display the constants **m0** and **m1** used in the curve fit and the equation they are for. Write all of this down.

4. Close the graph and return to the data. Go to the “Windows” menu and select formula entry. In the text box type, **$c2=m0*\exp(m1*c0)$** , substituting the numerical values for m0 and m1 that came from the curve fit. Then hit “Run”. This will fill in column 2 (c2) with the data generated by the curve fit equation, using the times in column 0. (c0)

5. Now, plot the actual data and the computer-generated data each on the y-axis, with the year again on the x-axis.

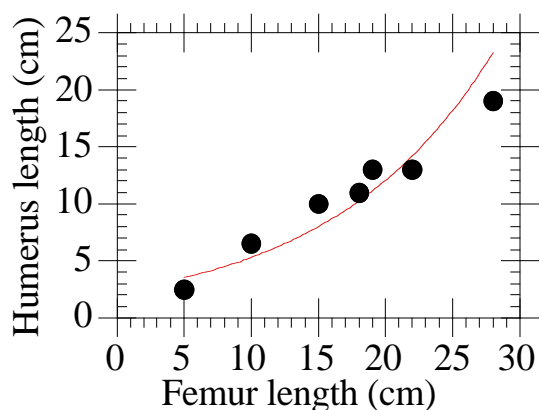
D How good is a fit? Correlation coefficients

To determine how well a set of data is fit by a given equation, you can use Kaleidagraph to find correlation constants (denoted as R). For a perfect fit, R will be 1.000. For a completely

incorrect fit, R will be 0.000. Correlation constants give a quantitative measure of how well a set of data is fit by a given equation and are therefore extremely useful to scientists. But beware--a high R value does not prove your fit is correct. R values are best used to *decide* between two different curve fits.

1. Go back to the Microsoft profits in the previous section, go to the “Curve Fit” menu, select “Exponential...” and at the dialog box click in the box below “View”. This will display the constants used in the curve fit. The correlation constant is listed as “R”. Write this down.
 2. Click on “Deselect” to get rid of the exponential curve fit.
 3. Fit the Microsoft profits using a linear curve fit, and write down the R value.
- Compare the R values for the exponential and linear fits. What can you conclude?

- Dr. Botchitt found that there was an exponential relationship between femur length and humerus length; to prove her assertion, she measured a correlation coefficient of 0.95 for the data shown. While Dr. Botchitt holds a press conference to announce that anyone over 6’ tall should have hands that go down to their knees, describe the major error in data analysis she made.



E. Post hoc, ergo proctor hoc

An additional issue that must be considered when analyzing data is whether a correlation arises as a result of a causal pathway or whether the correlation is merely coincidental. Even if a set of data correlates nicely, it is possible that the correlation is just a coincidence, or that both of the variables are in turn correlated to the same underlying cause. Correlation does not imply causality!! This error in thinking is so common, it has its own name (in Latin, yet): *post hoc, ergo proctor hoc*. So, scientists seeking to explain why a phenomenon occurs must a) demonstrate that there is a correlation and b) give some explanation of how the correlation is caused. The next exercise is intended to give you some practice with these two steps.

1. Go to the chem191 directory and open the file, “**U.S. pop vs. world car product**” which contains data for the number of cars produced in the world and the U.S. Population from 1950 to 1996. Plot this data, putting one variable (car production or population) on each axis. (It doesn’t

matter which variable goes on which axis) Then, fit a line to the data using the procedure described above.

2. Get the correlation constant for the curve fit, by going to the “Curve Fit” menu, selecting “Linear...” and clicking the “View” box next to the data set title. This will bring up a dialog box that includes the constants for the equation of the line as well as the correlation constant, R. Write down the value of R.

3. Repeat this procedure for two other files, “**College vs Prison**” and “**Presidents vs. Pig iron**”. “College vs Prison” contains data for the numbers of college students and prison inmates from 1975-1995. “Presidents vs. Pig iron” contains data for the number of letters in the president’s last name and the American output of pig iron from 1960 to 1995. Fit these data sets using a linear curve fit, and record all of the R values.

- Classify each of these three sets of data as correlated or uncorrelated. If the data are correlated, give some explanation of how the correlation is caused.

F Modeling World Population:

The goal of this part of the lab is to use the curve fitting techniques you have just learned along with a little math to generate a model for the population of the world from antiquity into the future. More details on the mathematical models we will use are found in Appendix A.

Different researchers come up with different values for world population, particularly for population in historical or pre-historical eras. It is not known which, if any, of these estimates are “correct,” so we will just take the mean values for different times and bear in mind that our data contain a potential source of error. We will assume that if any of the estimates are “way off” from the others, they are probably incorrect and should be removed. In order to determine which points are “way off”, you have to calculate the spread in the estimates, or the standard deviation, for each time period. Any population estimate that is more than 1.96 standard deviations away from the mean value will be thrown out, since it is outside the 95% confidence limits for the data.

1. Go to the chem191 directory and open the file, “**world pop data.**” This Kaleidagraph file includes population estimates from various sources from the year 10000 BC to the present.

2. Calculate the mean population value, over all the estimates for each year listed in the left-hand column. This would be rather nasty if you had to do it by hand, but thankfully, Kaleidagraph allows this to be easily automated. Go to the “Windows” menu and select “Formula Entry”. In the text box, type, **c12=mean([0:0,1:11])** and hit “Run”.

This command goes down each row of the table and calculates the mean by analyzing the cells in that row from columns 1 to 11. The mean is the average value of the population estimates.

3. Calculate the standard deviations over all the estimates for each year. Again, open the “Formula Entry” dialog box and type **c13=std([0:0,1:11])** and hit “Run”.

This command determines the standard deviation (S.D. or σ) of the data set for each year, based on the different estimates for that year. For years where there is only one population estimate, no standard deviation is calculated.

$$S.D. = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

where X_i represents each measurement, \bar{x} is the mean and n is the number of measurements.

Standard deviation is a measure of how spread out a set of data is. A high standard deviation indicates that the data vary widely, while a low standard deviation indicates that the data are more tightly spaced.

4. Points more than 2 SD from the mean are called outliers and can be discarded as statistically insignificant. Go through the data (the population estimates), and look for points that are more than 2 standard deviations away from the mean. (Points that have a value that is greater than the mean + 2 SD or less than the mean -2 SD) If you find any such points, they should be removed from the data set (by masking them--ask a TA how to do this) and the mean and standard deviation should be recalculated.

5. Plot the mean population (y-axis) (column 12) vs. Time (x-axis) to get a feel for how the world’s population had changed in the last 12,000 years.

- What kind of curve might fit this data? Try it and describe the result in your notebook. Record the correlation coefficient.

- Now expand the graph so you are just looking at modern times (say, 1900 to 2000) and comment on the goodness of the fit.

G.1 Specific Growth Rate and Human Population

Even though the entire data set may look like an exponential curve, the fit gets pretty bad in modern times. For an exponential curve, the rate of change of y divided by y (called the *specific growth rate*) is a constant. However, the specific growth rate for human population growth has not remained constant over time.

In an exponential equation, the *specific growth rate* m is defined as the rate of change of y (the derivative of y with respect to t , or dy/dt) divided by y .

$$\text{specific growth rate } m = (dy/dt)/y$$

This can be derived by taking the derivative of the exponential function $y = e^{mt}$:

$$dy/dt = me^{mt} = my$$

and dividing the left- and right-hand sides of the equation by y .

In the case of the world population the *specific growth rate* changes each year. So, when you try to fit an exponential curve to the world's population you can see that it grossly underestimates the current population because it is trying to find the best single *specific growth rate* to model the data when the actual *specific growth rate* is changing. But this does not mean that all hope is lost. If assumptions can be made about how the specific growth rate changes over time, then we can refine the exponential model to better fit our data. Three distinct historical trends will appear:

I. Constant specific growth rate: 10,000 BP-Industrial revolution. In this time period, human population followed an exponential growth formula.

$$y = y_0 e^{\mu t}$$

where y = population, t = time, y_0 = population at time 0, and μ = specific growth rate

II. Linearly increasing specific growth rate: Industrial Revolution-recent. In this time period, improvements in medicine, hygiene, agricultural production, and food distribution greatly increased average lifespan and caused superexponential population growth rates.

$$y = \frac{y_0}{1 - k y_0 \mu t} e^{\mu t}$$

where y = population, t = time, y_0 = population at time 0, and the specific growth rate $(dy/dt)/y$ is given by $\mu - ky$

III. Linearly decreasing specific growth rate: Recent times have seen a decreasing population

growth rate, caused by improving economic conditions and widespread use of birth control.

$$y = \frac{y_0 e^{\mu t}}{1 + \frac{M}{y_0} (e^{\mu t} - 1)}$$

where y = population, t = time, y_0 = population at time 0
and specific growth rate = $(dy/dt)/y = \mu - ky$, and $M = \mu/k$

In order to use these equations properly, you need to know whether the specific growth rate is increasing, decreasing, or remains constant over a given time period. Then, you can fit the data with the appropriate equation, and see if you get a good fit. The next sections will walk you through this process.

G.2 Finding specific growth rates during different historical periods

1) Create a new data worksheet by going to the “File” menu and selecting “New”. Then, copy the time and mean population data from “world pop data” to the new worksheet by clicking on the title cell of each column, selecting “Edit”:”Copy” and then clicking at the top of a column in the new worksheet and selecting “Edit”:”Paste”.

2) Determine the specific growth rate as a function of time. Mathematically, what is going on is this: you pick a time interval, say 10 years, and measure the change in population over that time interval. This gives you the growth rate dy/dt . Next, you divide by the population y in the start of the interval, and this gives you the specific growth rate at that time. You now repeat this process for another time interval. This would be quite tedious to do by hand, so let the computer do it for you:

3) Go to the “Windows” menu and select “Formula Entry”. Click in the bottom left corner of the box that is brought up (on the little picture of the note) and type **c2=(cell(index()+1,1)-cell(index(),1))/abs(cell(index()+1,0)-cell(index(),0))/cell(index(),1)** in the text space, hitting “OK” and then hitting “Run”. Disregard the alert pops up.

4) Next, plot *specific growth rate* against population. (*Specific growth rate* on the y-axis and population on the x-axis) Print out the graph and attach it into your notebook.

- Identify the time periods in which the specific growth rate is constant, linearly increasing, or linearly decreasing.

5) Once you have identified the years in which specific growth rate is constant, increasing, or decreasing, the next step is to extract constants to feed into the specific growth equations. Pick one time period and mask all the other data. To mask data, select it (so that it is highlighted), go to the “Functions” menu and select “Mask”. The selected data will turn red. Only unmasked data is available for calculations and for plotting. (Data can be unmasked by selecting it, then going to “Functions” and clicking on “Unmask”.)

6) Since the specific growth rate changes linearly with population, you can fit it to the equation of a line: $(dy/dt)/y = \mu - ky$. Go to the “Curve Fit” menu and select “Linear” to fit your specific growth rate vs. population data to a line. Once the best fit line has been plotted, go to the “Curve Fit...” menu, select “Linear” and check the “View” box next to “population”. This view box gives the constants used in the equation of the line.

- Write down the time period and the values for the constants which correspond to μ and k .

Now repeat steps 5) and 6) to find the constants μ and k for the other two time periods.

7) Unmask all the data.

G.3 Modeling the World population from antiquity into the future

1) The first step is to make sure there are no negative time values. Go to the "formula entry" box and type **$c_0=c_0+12000$** .

2) The next step is to use your values of μ and k to generate estimated population data for each of your three time periods. First, decide which equation you need to use in each time period (depending on whether you have an increasing, decreasing or constant growth rate).

3) Add three new empty columns to the data sheet by going to the “Data” menu and selecting “Append Column” three times.

4) Pick one of your time periods and mask all the other data. Generation of simulated population data requires you to a) get the correct equation for population growth and enter it in a form that Kaleidagraph can understand and b) include the proper values of μ and k for the time period. Sample equations for the linearly increasing and linearly decreasing growth rates, in Kaleidagraph format, are available in the Chem191.01 folder.

Go to the Windows menu and select “formula entry.” Click on the picture of a notepad in the

lower lefthand corner; this will pull up a notepad page. Now, click on “open” and select “linearly increasing equation”; substitute your own values of μ and k in this sample equation.

Example: for linearly increasing growth from 1900-2000, the appropriate equation is:
$$y(t) = [(y_0/\mu + ky_0) * e^{\mu t}] / [1 - (y_0/\mu + ky_0) * k e^{\mu t}]$$

From the data, the initial population was 252. If the measured value of μ is 0.000215 and k is 0.0000822, then the equation that should be entered into Kaleidagraph is:

$$c3 = ((252/(0.000215 + 0.0000822*252))*0.000215*\exp(0.000215*(c0-1900)))/ (1-(252/(0.000215 + 0.0000822*252))*0.0000822*\exp(0.000215*(c0-1900)))$$

Note that the times (from column c0) were adjusted (time is entered as “c0-1900”) so that the equation runs over the interval 0-100 rather than 1900-2000.

Once the formula has been entered, hit “OK” and then “Run”. This plots the data from the equation in the empty column, column 3, for the time values available.

7. Plot time against the actual population and the population estimates of your model.

- How good does the fit look? If you project the population estimates into the future, will this model overestimate or underestimate the population?

8. Open the file “U.S. Population” from the chem191 directory. This file contains actual U.S. population data for this century. Plot the population against time.

9. Unlike the case with the population of Exampilia, the U.S. population can be characterized by a constantly decreasing *specific growth rate*. Eventually, such a population will reach a stabilization point, as the *specific growth rate* decreases to zero.

- When the specific growth rate is zero, what is the population? Remember, the specific growth rate = $\mu - ky$, so if that equals zero then y , the population, can be expressed in terms of μ and k .

10. Calculate the specific growth rate for the US population data.

11. Plot *specific growth rate* against population. (*Specific growth rate* on the y-axis, population on the x-axis) Go to the “Curve Fit” menu and select “Linear” to fit your specific growth rate vs. population data to a line. Once the best fit line has been plotted, go to the “Curve Fit...” menu, select “Linear” and check the “View” box next to “population”. This view box gives the constants used in the equation of the line.

- Write down the values for the constants which correspond to μ and k .

12. Go back to the data sheet and if necessary add a new empty column by going to the “Data”

menu and selecting “Insert Column”. Now you need to get the correct equation and enter it in a form that Kaleidagraph can understand. A sample equation for linearly decreasing growth rate is available as “linearly decreasing equation” in the Chem191 folder. Go to the Windows menu and select “formula entry.” Click on the picture of a notepad in the lower lefthand corner; this will pull up a notepad page. Now, click on “open” and select “linearly decreasing equation”; substitute your own values of μ and k in this sample equation.

Example: for linearly decreasing growth from 1900-2000, the appropriate equation is:
 $y(t)=M/[1+ e^{-\mu t}((M/y_0)-1)]$ (Where $M = -\mu/k$)

From the data, the initial population was 76.09. If the measured value of μ is 0.0794 and k is -0.0001323, then $M=600$ and the equation to enter into Kaleidagraph becomes

$$c3 = 600/(1+(\exp(-0.0794*(c0-1900)))*((600/76.09)-1))$$

Once the formula has been entered, hit “OK” and then “Run”. This plots the data from the equation in the empty column, column 3, for the time values available.

13. Plot time (x-axis) against the actual population and the population estimates of your model.

- How good does the fit look?

14. Add some extra time points to $c0$.

You can do this by hand by typing in 2000, 2025, etc, or you can click on $c0$ and then go to the “Functions” menu and select “Create Series”. For the initial value put 1900, the increment could be 5, the multiplier 1, and the final value 3000. Click on “OK” to create the series.

15. Now, rerun the formula to generate the predicted population of the US up to the year 3000.

- In which year does the population stabilize? What is the predicted stable population?
- Compare the predictions of an exponential model with this one. You may find it helpful to do an exponential curve fit to the real population data. In what year would we know which model is “right”?

Writing up your lab report

Make sure you have answered the various questions throughout the lab in your lab notebook.