Scientific Data

What do I do with it?

and

What is it telling me?

Scott A. Sinex
Barbara A. Gage
Department of Physical Sciences
Prince George's Community College

2001

An Excel spreadsheet and STELLA model accompany this activity.

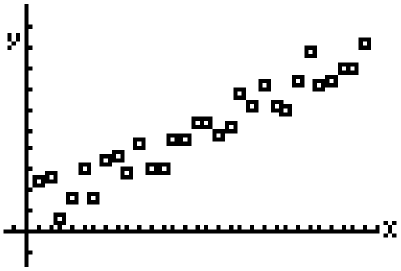
To understand simple and complex systems, scientists collect data from a variety of laboratory experiments and environmental measurements from instruments such as sensors on satellites. This data may reveal relationships that explain the behavior of the system under study. How do they handle the data? To provide you with an understanding, we will provide you with a chance to explore data handling on a smaller scale in this activity.

Here are some of the processes we need to undertake when given a set of data:

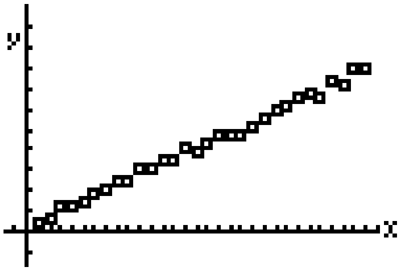
❐ produce a graph of the two variables, usually a scatter plot;
❐ look for trends in the data, describe them in a qualitative fashion;
❐ find a mathematical relationship to describe the trend, and;
❐ answer the question “Does the trend really mean something?”

When we plot data, we want to know if there is a relationship between the x and y variables. Look at the six x-y plots given below and address this question for each: As the x-variable changes, how does the y-variable behave?

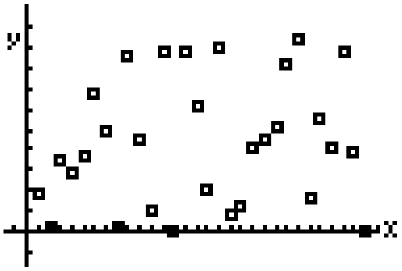
A



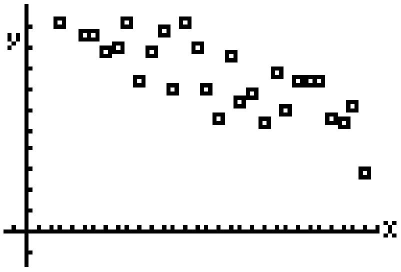
B



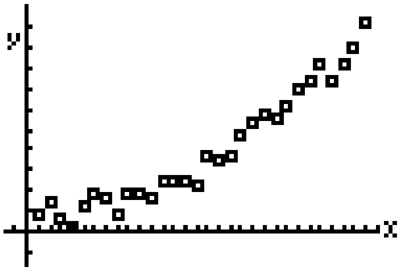
C



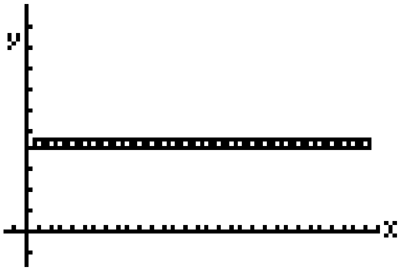
D



E



F



In looking at the behavior of the variables, you are trying to decide if a trend exists in the data. Which plots show a trend?

Now let's investigate some unplotted data seta. The data sets you will be working with are set up in Excel such that each set is a separate worksheet in the same file. The file is called "Scientific_Data" and is found at http://academic.pg.cc.md.us/psc. When you get to the site, click on student resources. Graphing can be done by hand on graph paper or with the graphing calculator or a computer spreadsheet such as Excel or Vernier's Graphical Analysis. If you use Excel, the graphs you generate for each data set can be placed on the worksheet for that set. Instructions for the use of the graphing calculator, Excel, and Graphical Analysis are on the Physical Sciences Department web page at http://academic.pg.cc.md.us/psc under student resources.

Data Set #1 – Ozone

Let's examine a simple data set of the annual average levels of ozone measured over the years 1956 to 2000 at Halley Station in Antarctica. Generate a scatter plot of this data. Time in years will be the independent variable, so it will be plotted on which axis?

Sketch the graph and label the axes.

How would you describe the behavior of ozone concentration over time?

What might you do with a straight edge on the graph to help explain the behavior of ozone over time?

Is there a mathematical relationship that could describe the data?

The graph shows a decrease in ozone over the 40-year period of measurement. The trend looks linear and a straight line could be drawn through the data using a straight edge. A better method to visualize the trend would be to do a linear regression, which is an option available on the graphing calculator, Excel (called a trendline) or Graphical Analysis. Linear regression is a data-fitting technique that draws a line by minimizing the difference between the data points and the best-fit line. Using regression fits of data also allows us to examine the goodness of fit using the coefficient of determination, r^2. We will address this later.

Data Set #2 – Carbon Dioxide
Our next set of data is the annual level of carbon dioxide measured in air at Mauna Loa Observatory in Hawaii from 1970 to 1999. Now plot a graph of the two variables found in the data table. The independent variable is placed on the x-axis, while the dependent variable is placed on the y-axis. The dependent variable changes due to the change in the independent variable. Does the carbon dioxide depend on time or does time depend on the carbon dioxide? What is the dependent variable in this case?

Sketch the graph and label the axes.

Is there a trend? If so, describe the trend.

Can you determine a mathematical relationship that describes in equation form how the two variables are related? Write the equation in terms of the variables studied (NOT x and y).

How well does the relationship fit the data? Explain.

How does the trend of the ozone data compare to the carbon dioxide data? List any similarities and differences.

Data Set #3 – Dissolved Oxygen
This data set is the dissolved oxygen level in stream water over a variety of temperatures for one year at Shepherdstown on upper Potomac River. What is the independent variable?

Is there a relationship between temperature and dissolved oxygen?

Derive a mathematical model for the relation above if it exists.

Data Set #4 – Temperature

This set shows the water temperature of Jack Bay on the Patuxent River over three years time or 36 months. The data starts in January 1998 through December 2000.

Plot the data and describe any trend.

What is different about this data compared to any of the earlier sets?

Data Set #5 – Atmospheric Pressure

Here is a data set for the variation of atmospheric pressure with altitude. Plot a graph of atmospheric pressure against altitude.

How would you describe the trend?

Commercial airlines fly at a maximum altitude of 11 km, why can't you open a window at this altitude on an airplane?

Modeling data is a process that consists of four steps:

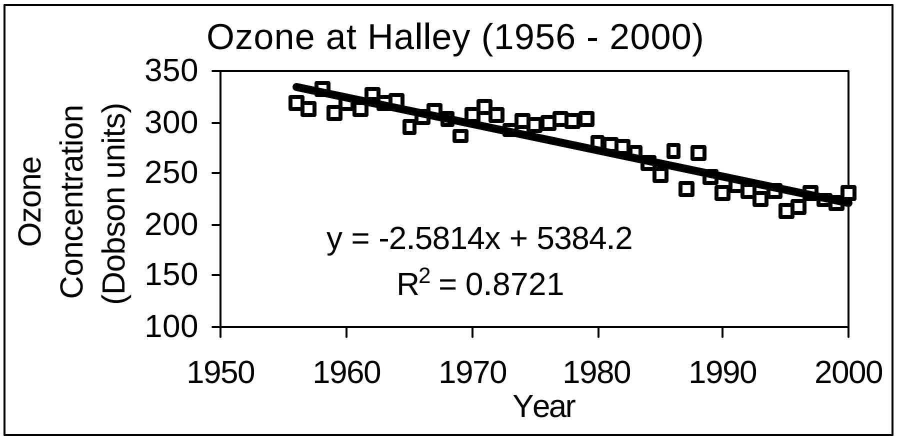
1. Generate a scatter plot of the data collected.

2. Pose the question: Is there a significant relationship between the two variables plotted?

3. Determine how well the mathematical relationship fits the actual data.

4. Use the model to make predictions. Do the predicted results make sense or does the model break down?

Let's examine the ozone data for Halley Station again. Below is the graph with a linear regression displayed with the data points. The data points are scattered around the regression line. For the number of data points involved, the relationship is statistically significant (for the r^2 value).



Now let's translate the equation into terms of the variables studied from the regression equation.

Ozone concentration = - 2.58 (year) + 5.38 x 10^3

This is an important step to complete, as very few science books ever use x and y in expressing mathematical relationships. Look in your textbook and see!

According to the model, what will the ozone concentration be in 2010?

At the present rate of ozone removal at Halley Station, when will the level drop to zero? Explain how you determined this.

Do you think is this a real possibility? Explain.

How do you judge the goodness of fit?

When a line is fit to data by a regression technique, the goal is to minimize the difference between each datum point and the line. The difference is called a deviation. Regression techniques minimize the squared deviations to find the best-fit line or curve. The coefficient of determination, r^2, is an overall measure of how well the data and the regression line agree. We ask the questions "Is the data on the line or scattered about the line? " and "If it is scattered, how do you judge how far is too far away from the best-fit line before the relationship is not significant?"

The value of the coefficient of determination, r^2, tells you the fraction (or multipled by 100, the percentage) of the y-variable change explained by the change in the x-variable. A value of $r^2 = 1$ would mean that 100% of the y-variable change is explained by the x-variable change.

On the worksheet labeled r-squared is an Excel interactive spreadsheet that will allow you to copy and paste four different data sets one at a time and judge their goodness of fit. Analyze each set and comment on the goodness of fit for each data set.

Data Set	r^2 value	Comments
A		
B		
C		
D		

Best fit data set: ________ Worst fit data set: ________

Data Set #6 – Aluminum Cans
Develop a mathematical model to examine how the mass of an aluminum can has changed over the last 30 years. Follow the steps for developing a model given in this activity.

Write the mathematical equation that describes your model in term of the variables studied (NOT x and y).

What are the units of the slope?

What does the slope mean in terms of the model?

How good a fit is your model? Explain.

If the model above is correct, in what year will the mass of a can go to zero?

Can the mass of an aluminum can go to zero? What does this say about the present model?

Models and Simulations

We have seen that a model can be a mathematical relationship that describes the behavior of a system. We have been developing two variable (x and y) models but other variables may enter the mathematical equation on further study. We can eliminate them as variables by holding them constant during an experiment. You will see this later in the semester when you investigate the gas laws and spectrophotometry.

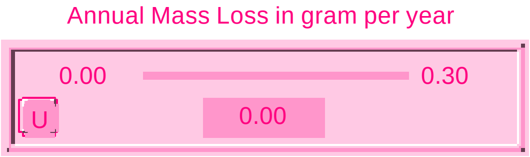
Once a model is developed we can test it by seeing how it behaves when we change a variable and/or extend the numerical range of the variables. This is referred to as a simulation.

We have a STELLA model of the aluminum can mass loss model you developed above. If you are going to download the STELLA model in a college computer lab **use Internet Explorer and save the model to the desktop.**

On a college computer in a lab Click on the ACADEMIC APPLICATIONS folder, then the MATH folder and find STELLA™ and boot it up. Close the new model screen (inner screen), go to file and click on open. Go to the desktop and select the model.

If you have downloaded STELLA™ on your home computer and have loaded it on your hard-drive Click on the model you want. This will open the model to the simulation page, which is where you want to be. (If this does not work, boot STELLA first, then close the new model screen (inner screen), and then go to file and then click on open and get to the A: drive or where ever you placed the model and select the appropriate model.) You will download version 7, which will run the models (version 5).

What is the annual loss of aluminum from a can from your model? ______ Set this number on the slider in the STELLA model, which is illustrated below.



What does the model predict for the mass of a can in 2012?

The aluminum can industry feels that the can will get to a mass of 12 g in the future. Using the pause function on the model, what year is this going to happen based on the present model?

The linear model for the mass loss of aluminum cannot go on forever or to zero. There has to be a limiting mass of aluminum. The linear model breaks down at this limiting mass. Sketch a graph of how this model must respond to the limiting mass.