Introduction to Graphs and Linear Regression

Content Discussion and Activities

PHYS 104L

1 Goal

The goal of this week's activities is to provide a foundational understanding regarding using graphs to represent data. In addition, a familiarity with "best fit lines" and a Linear Regression calculation will be developed. The importance of labeling, units, and uncertainties as well as using Excel to complete expected graphing work are primary topics to be covered.

2 Introduction

Making a graph of data or values calculated from data is one of the most useful things done by scientists. Having a visual representation of data can help to give deeper insight into the way the plotted data relate to each other, and therefore this has the potential to help us understand something about the physical system under consideration. In this lab you will take some data and plot it. We will review some basic graphing ideas and then talk about some basic graphical analysis. In addition, we will discuss the idea of a "best fit line" to graphed data and the use of a Linear Regression calculation to determine the "best fit line" using software. We hope that you will learn how to make graphs in a correct way, understand something about what a graph can tell you about the variables plotted, and also understand how to determine the linear equation relating quantities expected to be linearly related.

3A Discussion - Graphing Data

The ability to use a graph, either to determine, display, or understand data and relationships between physical quantities is a primary skill for study in the sciences. You have probably been using graphs since your days in elementary school and have even constructed several graphs in different classes prior to now. This section of the handout deals with constructing a graph if given data for two different quantities.

Typically, when studying a system, a scientist is trying to determine how one quantity depends on other quantities. In an ideal situation, an experimenter would be able to hold all the possible variables that could affect the quantity of interest constant except for one variable of interest. In this case, as the value for this one quantity is varied, the value of the quantity of interest can be measured. The quantity which was being varied, or set, to different values is called the **independent variable,** and the quantity of interest, whose value was measured once the independent variable's value was set, is called the **dependent variable**. The output from this process would be a set of corresponding values for the independent and dependent variables. By graphing these values, the relationship between these two variables can be studied. The simplest graph we can make is a scatter plot on a Cartesian Coordinate system with the variable values plotted as (x,y) values on the graph.

Unless there is a good reason (related to the analysis of the data) otherwise, usually the independent variable is plotted along the horizontal axis (used as the x values) and the dependent variable is plotted along the vertical axis (used as the y values). Later this semester, we may come across some cases where we will want to plot the data with the independent variable on the vertical axis, but such is not the norm. Graphs can be drawn and sketched by hand or via use of software graphing programs. For graphs to be useful, attention must be paid to maintaining a consistent scale along both horizontal (x) and vertical (y) axes. This consistency is more difficult to maintain when making a hand sketched graph, however, the use of pre-gridded graphing paper makes doing so easier. Whenever possible, we will want to use the computer for making graphs and performing analysis on them. In that case, the computer will do much of the difficult work for us, leaving us to only have to interpret and understand the graph. Understanding what goes into making a graph, however, will help us understand and read graphs the computer makes for us.

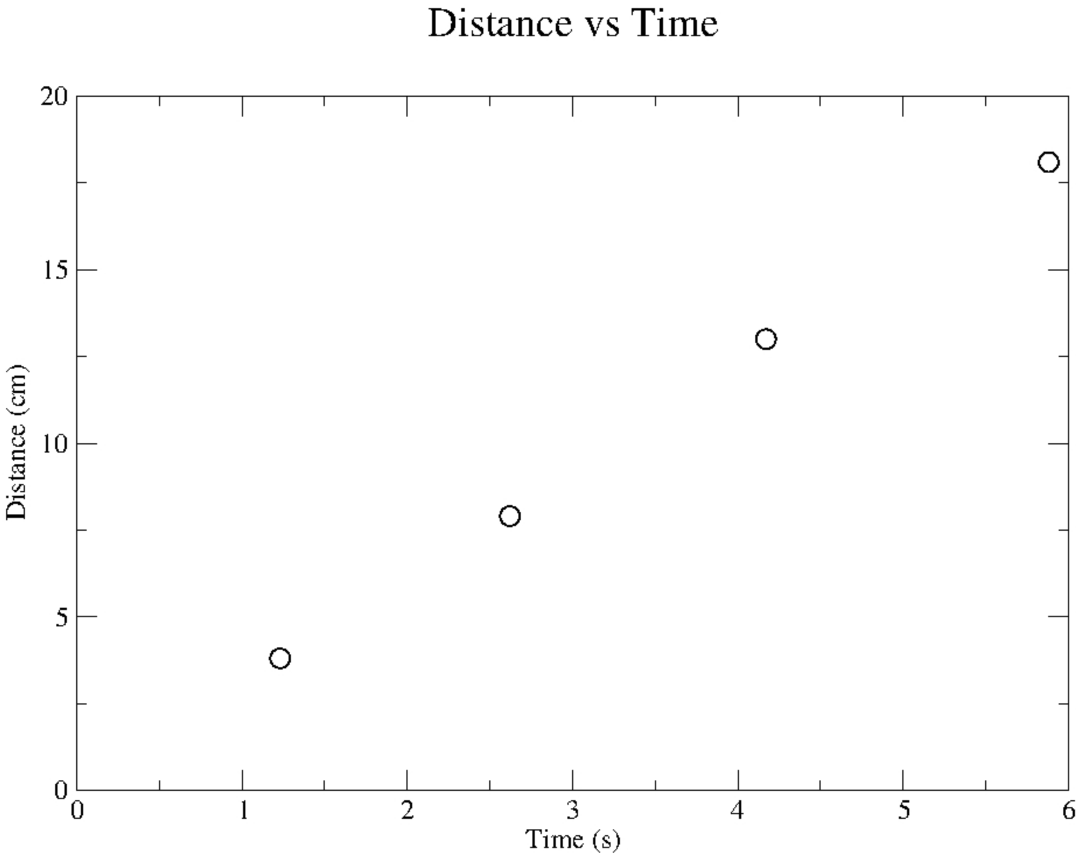
Let's assume you have several sets of corresponding values for the independent and dependent variables (x,y) and are ready to construct a graph. You have already determined which variables will be plotted on the horizontal and vertical axes so the next step is to identify the scale and range for those axes. Look at the x values for your data points. Identify the highest and lowest values you will use. This is the range of values your horizontal axis must include. You will want to start the x axis at a convenient value somewhat lower than your smallest x value (or at least equal to the lowest x value) and have the horizontal axis end at a value slightly larger than the largest x value (or at least equal to it). After identifying a convenient increment to break that range up into, one can then label the x axis. For example, if the x values were going to range from 5 to 67, I might decide to have the horizontal axis of my graph start at 0 and go up to 70 in 10 unit increments. The same process should be followed for the vertical axis based on the y values to be plotted. Once the axes are labeled, the actual data points can be plotted as (x,y) Cartesian points on the graph. One should always be sure to label what is plotted on each axis, both the quantity and what units that quantity is measured in. The axis should also have numerical values on it. Often , the graph axis is labeled with "tick marks" to indicate equal increments along the axis. You will also want to label the graph with a title. The typical convention is to describe a graph as " 'y-axis quantity' vs 'x-axis quantity' ". For example, a graph with total rainfall on the vertical axis and month of the year on the horizontal would be labeled as a graph of Total Rainfall vs Month of Year.

As an example, assume you have timed an object that moves in a straight line in the x direction. You could collect the data by starting the stopwatch when the object passed a certain point, and then looking to see where the object was after a given amount of time had passed. In this manner, the elapsed time would be the independent variable and the distance travelled would be the dependent variable. The distance traveled depended on the amount of time you let go by before measuring where the object was. You can repeat the measurement at a given distance any number of times that you like, and you can also vary the elapsed time that the object moves for a given trial.

You end up with a data table that looks something like this:

Time (s)	Distance (cm)
1.23	3.8
2.62	7.9
4.17	13.0
5.88	18.1

If you then plot distance versus time, you get a graph that looks like this:



Notice the labeling for the axes. Does it make sense how the four data points were plotted?

3B Activity – Constructing a Handmade Graph

1.) The table below gives data (fake) for number of kids at a pool for different daily temperatures. You are to make a graph of Number of Kids vs Temperature by hand using the available graphing paper. Be sure and label your axes with units and at least two numerical values on each axis.

Daily Temperature (Fahrenheit)	Number of Kids at the Pool
71	140
75	150
81	175
85	210
88	250
92	335
96	420
104	550

2.) Here is a second set of made up data to consider. The listed ordered pairs give the number of hours spent studying by different students for the Math SAT and their resulting Math SAT score (hours studying, Math SAT score). Make a graph of Math SAT score vs Hours Studying appropriately labeling the graph as expected.

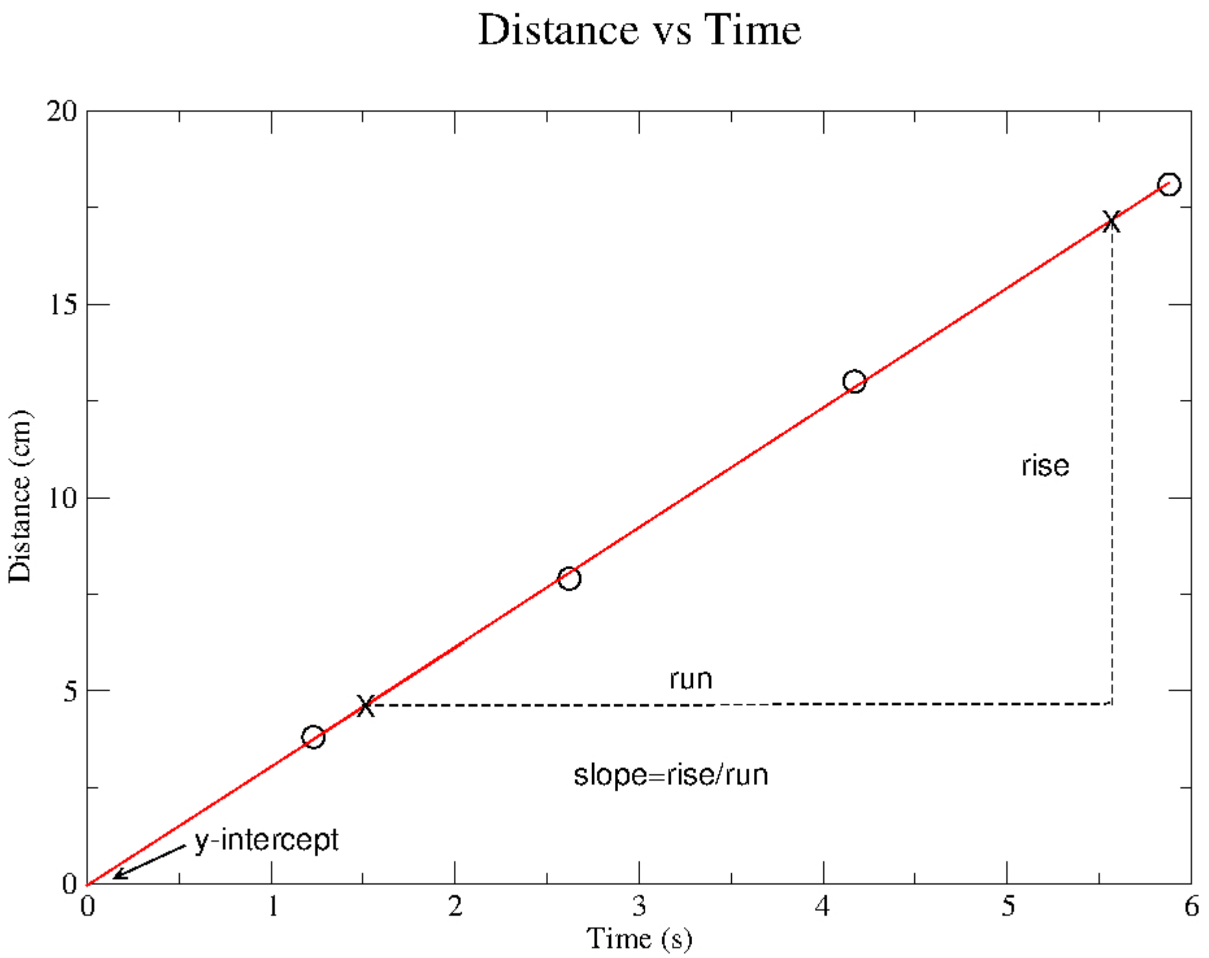
(4, 395) (13, 720) (9, 570) (10, 640) (7, 520) (4, 390) (14, 780) (9, 600)

(12, 630) (2, 340) (5, 450) (6, 530) (11, 680) (10, 690) (16, 760)

4A Discussion – "Best fit line" for linear data

If we look at the two graphs from the previous activity, we can see that they have different shapes. The first graph is curved while in the second graph the data points scatter along a relatively straight line path. The second graph has data which appears to be **linear**.

At some point in your past mathematical life you learned the equation for a straight line: y=mx+b . In this equation, y represents whatever variable is plotted vertically, and x represents whatever is plotted horizontally. In such a case, we say we have graphed " y versus x " when we refer to the graph. Our prior example graph (above) of Distance vs Time appears to be very linear.



When we have a graph which we expect to be linear, we will often refer to the “best fit line” to the data. The “best fit line” is the straight line that comes closest to going through all the data points on the graph. If we wanted to determine the best fit line on this graph, we would take our ruler and adjust it until we thought we had the best representation of the data; the data points would scatter around the line as close to it as possible. Then we would draw the line on the graph. In an ideal situation this line would go through every point on the graph, and things would look perfectly linear. In practice our line will not go through all of the points, but the line we draw should come as close as possible to as many points as possible. When we draw this line, we also might want to make sure that it extends through the y-axis. This, then, would allow us to better estimate the y-intercept of the line that we drew on the graph. To find the slope of this line, we should pick two points on the line, fairly widely spaced from each other, and then figure out the rise and the run between these two points. The ratio of the rise to the run is the slope of the line, as shown in the figure reconsidering the Distance vs Time graph.

4B Activity – Best fit line

1.) Go back to your second graph from the first activity. Using your ruler, draw in your best fit line to the data. Make sure this line extends across the entire graph. Use this best fit line and your graph to determine the slope (m = rise/run) and the y-intercept (b) for this best fit line. Record these values on your activity sheet and show how you determined the slope. Remember, the slope and y-intercept have units that go with them. The units of the y-intercept will always be the same as the units for the y-axis. The units for the slope will always be the ratio of the y-axis units to the x-axis units.

5A Discussion – Using Excel to graph and perform Linear Regression Calc.

You will have now made two hand constructed graphs and drawn in the best fit line for the graph that looked fairly linear. While hand drawn graphs may allow us to get a rough idea of how the quantities are related to each other, the usefulness of the graph depends on the care we took with plotting points. Using a computer for constructing our graphs will give more accurate representations of the data and likely be done quicker. In addition, most graphing programs have the ability to perform a **Linear Regression** calculation for a set of data. A Linear Regression calculation uses an algorithm to determine the slope and y-intercept of the best fit line. For determining the best fit line, the calculation identifies the line which minimizes the sum of the distances between the data points and the best fit line. You will be given a handout for using excel to make a graph and complete a linear regression calculation and your instructor will go over the process. For the rest of the semester, if you are asked to construct a graph, you should use Excel (or some other graphing program if you choose) to make your graphs and determine the equation of the best fit line if needed.

5B Activity – Using Excel to make graphs and find best fit lines.

1.) Use Excel to make a graph of Math SAT Score vs Hours Studying using the data from the first activity again. Complete a linear regression calculation and format the graph as shown in the example graph with the excel handout. For expected values, use the slope and y-intercept you determined from your hand constructed graph from activity 2. You will have two four line summaries under the graph, just as in the example graph, comparing the result from your linear regression analysis with the expected values for both the slope and y-intercept. Print a copy of this graph (it should all be on one page) to attach to your activity sheet.

6A Discussion – Physical Meaning of Slope or Y-Intercept

Sometimes, the reason we make a graph is to try and identify the relationship between two quantities. More often, we believe we know what that relationship is and are either making the graph to test that theory or are going to use the graph to determine the value for a quantity of interest. In these cases, we can use the expected or known relationship to either identify expected values for the slope and y-intercept or to identify the physical quantities the slope and or y-intercept of the graph should correspond to. As an example, assume you have a container with several, 30 or so, similar screws. If you were to take some number, N, of those screws and measure their combined mass (M), you might expect that the combined mass would be related to the number of screws you picked up. If we let n = the average mass of a single screw, then we would expect that

$$M = n * N.$$

We could envision collecting data by picking different values for N and measuring the combined mass M for each. If we were to construct a graph of M vs N, what would we expect that graph to look like? Recall the equation of a straight line is "y" = m "x" + b, where m and b, the slope and y-intercept of the line, are constants. For our graph, "y" is just M (the combined mass) and "x" is N (the number of screws measured). Comparing the straight line equation with the expected relationship between M and N, we can see that the constant n is the expected slope of the best fit line and the expected y-intercept would be 0.

$$\begin{array}{ccccccc} M & = & n & N & + & 0 \\ \text{“y”} & = & m & \text{“x”} & + & b \end{array}$$

If we were to perform a linear regression with this data, we would expect the y-intercept result to be in agreement with zero and the slope result would tell us the average mass of a single screw.

6B Activity – Physical Meaning of Slope and Y-Intercept

1.) You should have a container with 30 or so screws in it. You will collect data by measuring the combined mass of 12 different combinations of screws ranging from 3 screws up to 25. Record both N and M in the table on your activity sheet. Pick values for N that are spread out over the range of 3 to 25. After taking data for each combination, you should put all the screws back into the container and then pick out the next number at random, rather than just adding more screws to the ones you are already using.

2.) Pick three different screws at random and measure their individual mass. Average these values to determine your estimate for the average mass of a single screw.

3.) Use Excel to make a graph of M vs N and perform a linear regression. Compare the slope and y-intercept from the linear regression with expected values based on the given relationship between M and N.

4.) Did your results from the graph agree with your expected values? Briefly discuss what this outcome tells you. Do this below the 4 line summaries on your graph page.

Introduction to Graphs and Linear Regression

Activity Data Sheet

PHYS 104L

Activity 4B

Determined Best Fit Line Slope ____________________ Y-Intercept ____________________

Work for determining Slope:

Activity 6B

Number of Screws Used	Combined Mass (grams)

Mass of 3 Individual Screws ________________ ________________ ________________

Average Mass of a Screw ______________________