

Linear Regression

Science is a search for an understanding of the physical and biological universe. However, some of the things we would like to investigate are too large, too small, too expensive, or too dangerous to have in our laboratories. In these cases, scientists and engineers turn to *models*. Models are a representation of the real thing but of a reasonable size and cost. By studying the model, we can make some predictions about the behavior of the real thing.

Physical models are often used in engineering. For example, models of the space shuttle were flown many times in wind tunnels before the shuttle's design was finalized and the shuttle was constructed. Architects construct cardboard houses to help homebuilders visualize their home before construction begins. Molecular models help students and chemists understand the interaction between molecules and are often used to design therapeutic drugs.

Behind each of these models is a set of measured numbers from which the model is developed and from which predictions of the model's behavior are made. An equation is a mathematical model, a statement of a perceived relationship between variables. The advantage of mathematical models is the ability they provide to extrapolate and interpolate measured data. One of the easiest ways to develop a mathematical model involves graphing experimental data.

Using Graphs to Identify Mathematical Relationships

Simple X-Y graphs are constructed on a two-coordinate axis. The horizontal (X) axis is the **independent** variable. We set this variable when we make the measurement. The vertical (Y) axis is the **dependent** variable. This is the number we measure after we set the X value. When the plotted data points form a straight line, a mathematical relationship exists between the two variables that is represented by the algebraic equation $y = mx + b$ where y is the vertical axis value, x is the horizontal axis value, m is the slope of the line, and b is the y-intercept.

The graphing calculator performs a **linear regression** analysis by plotting the "best fit" line through the data and then writing the slope-intercept equation relating the two variables. Linear regression is also referred to as "least square" curve fitting. The "best fit" line is drawn through the data, the distance between each point and the line are determined, and these distances are squared. One then adjusts the slope and position of the line until the total of all the squared deviations from the line are minimized; thus the method of "least squares." Fortunately, a graphing calculator will do this for you.

The **correlation coefficient** is a measure of how well the regression line fits with the observed data. A perfect fit produces a correlation coefficient of either +1.000 or -1.000, depending on if the line slopes up or down. The closer the correlation coefficient is to +/-1.000, the better the regression line expresses the relationship between the two variables.

Note that many cause and effect relationships are not linear, and other types of regression analysis can be attempted (including power, logarithmic, etc.) We will focus exclusively on linear relationships in this lab since they appear quite often in general chemistry courses.

Procedural Notes: Complete problems 1-5 below using your calculator's linear regression abilities to answer the questions. Each linear regression must have a calculated correlation coefficient r (and *not* R^2 !) Graphs should be done in Excel (or a similar graphing program). Also complete question five which ensures you are subscribed to the **mhchem** mailing list. There is no need for a purpose or conclusion.

Problem 1: The Relationship Between Celsius and Fahrenheit

In 1724, the German scientist Gabriel Fahrenheit developed a temperature scale based on phenomenon he thought could be easily repeated in laboratories around the world. For his zero degree point, Fahrenheit chose the coldest mixture of ice, water, and salt that he could produce in his laboratory. For ninety-six degrees, he chose what he believed to be normal body temperature. Fahrenheit wanted a temperature scale that could be divided into twelfths. On this scale, pure water freezes at 32 degrees, and pure water boils at sea level at 212 degrees.

A few years later, in 1742, the Swedish scientist Anders Celsius developed a different temperature scale. This scale used pure water as its standard. Zero degrees was the temperature where pure water froze, and one hundred degrees was the temperature where pure water boiled at sea level. Because Celsius had one hundred degrees between the two reference points on his temperature scale, it was called the *centigrade* scale. Recently this was renamed the Celsius scale in honor of Anders Celsius.

The *Kelvin* scale was adopted as the temperature scale for the *Systeme Inetrnationale*, and the closely related Celsius scale has great merit in the laboratory and in everyday existence. As a result, temperatures measured in the United States using the Fahrenheit system must be converted to the Celsius scale to be meaningful elsewhere in the world.

A student measures the following data points in the laboratory using two thermometers:

Temperature (°C)	20.0	40.0	60.0	80.0	100.0
Temperature (°F)	67.6	104.8	141.1	175.0	211.1

1. Construct a graph of degrees Fahrenheit (y) as a function of temperature in degrees Celsius (x). Which variable is independent? Which is dependent?
2. Using your calculator, determine the mathematical equation of °F as a function of °C as well as the correlation coefficient, r . Record r to five significant figures.
3. Using the actual equation $^{\circ}\text{F} = 1.8^{\circ}\text{C} + 32$ and your experimental equation, convert 29.0 °C to °F. Calculate percent error = (difference / average) x 100% Comment on discrepancies.

Problem 2: Solubility of Lead Nitrate in Water

The solubility of lead nitrate in water was measured as a function of temperature. The solubility is given in units of grams of lead nitrate per 100 grams of water.

Temperature (°C)	20.0	40.0	60.0	80.0	100.0
Solubility (g / 100 g water)	56.9	74.5	93.4	114.1	131.1

1. Graph the data. Which variable is independent and which is dependent?
2. Determine the equation of the best fit line. Record the equation and correlation coefficient.
3. What is the solubility of lead nitrate at 47.0 °C?

Problem 3: Colorimetry

The colors in the visible spectrum of light are shown by a rainbow. Colored substances absorb segments of the visible spectrum of light. Pink solutions, for example, are pink because they absorb green light and transmit all other colors of the visible spectrum. If light of the particular color absorbed is passed through a sample, the amount of light absorbed will be related to the number of absorbing molecules in the light beam. Dilute solutions absorb little light, concentrated solutions absorb more. Typically the amount of light transmitted through the solution is measured; *transmittance* is inversely proportional to *absorbance*. The following data was obtained for the transmittance of 525 nm light by solutions containing different concentrations of permanganate ion.

Concentration (mg/100 mL)	1.00	2.00	3.00	4.00
Transmittance (unitless)	0.418	0.149	0.058	0.0260

1. Convert the Transmittance values to Absorbance using the following equation: $A = \log(1/T)$, where A = Absorbance and T = Transmittance.
2. Graph the Absorbance (y) versus Concentration (x).
3. Perform a linear regression analysis. Record the equation and the correlation coefficient.
4. Predict the absorbance of 2.50 mg permanganate ion / 100 mL solution.

Problem 4: Kinetics

The branch of chemistry that studies the rate or speed of reactions is called *kinetics*. One must often plot concentration versus time data in a variety of mathematical formats to find a linear relationship; this assists in finding the *order of reaction*. We shall explore this topic more in CH 222. The following data was collected at 25.6 °C while measuring the disappearance of NH₃:

Concentration [NH₃] (mol/L)	8.00×10^{-7}	6.75×10^{-7}	5.84×10^{-7}	5.15×10^{-7}
Time (h)	0	25.0	50.0	75.0

1. Prepare a graph of $\ln[\text{NH}_3]$ versus time. "ln" stands for natural logarithm which can be calculated easily on your calculator (for example, the value of 8.00×10^{-7} is -14.039.) Perform a linear regression analysis on the $\ln[\text{NH}_3]$ versus time data and find the equation and the correlation coefficient.
2. Prepare a graph of $1/[\text{NH}_3]$ versus time (for example, $1/8.00 \times 10^{-7}$ is 1.25×10^6). Perform a linear regression analysis and find the correlation coefficient and the values for the slope and the y-intercept.
3. Which graph gives a *better* linear regression? Why?
4. Plots of $\ln[\text{NH}_3]$ versus time that are linear are called *first order reactions* while graphs of $1/[\text{NH}_3]$ versus time that are linear are called *second order reactions*. What order of reaction does the decomposition of NH₃ follow?

Problem #5 on Next Page

Problem 5: Join the mhchem mailing list!

mhchem is a mailing list for my students in Chemistry 221, Chemistry 222 or Chemistry 223 at Mt. Hood Community College. Over the course of the year, you will receive various pieces of information, events, announcements, and more via mhchem, including at least one extra-credit assignment. The goal of this section is to *subscribe* (i.e. join) mhchem.

Send your Web browser to the following address:

<http://mhchem.org/mhchem>

Next, enter your **email address** and **name** in the appropriate places (if you do not have an email address, contact the instructor immediately to get one.) Press the **Subscribe** button; you should now be a part of mhchem.

For this lab report, **include the email address you used to subscribe yourself to mhchem under question #5** to receive credit. Simple, no? ☺