Using Linear Regression to Determine Plate Motions

Michelle Kathleen Hall-Wallace
Department of Geosciences
University of Arizona
1040 E 4th Street
Tucson, Arizona 85721-0077
hall@geo.arizona.edu

ABSTRACT

Scientists commonly gather data and develop equations to describe relationships among data and variables using linear regression. Providing geoscience majors opportunities to determine physical relationships using regression techniques is important for their understanding of the nature of science. Fortunately, regressions are easily calculated with spreadsheet or statistics software, and, if the mathematical basis is well developed, students can understand the predictive power of a regression and apply it to many problems.

In the activity presented here, undergraduate geoscience majors use linear regression techniques to determine rates of Pacific-plate motion over the Hawaiian hotspot through time. Using age and location data for the Hawaiian-Emperor volcanic chain, students calculate the rate of plate motion for the entire chain and the separate components, then determine whether plate motion has been constant over time. Using latitude and longitude data, they determine the location of the bend in the volcanic chain. Finally, they develop a relationship between age and location to make predictions about where existing volcanoes will lie in the future and the age of the bend in the volcanic chains. Students are introduced to error analysis by examining data errors and learning about the sources of those errors and by evaluating formal errors calculated in the regression analysis.

Keywords: Education – computer assisted; education – geoscience; education – undergraduate; miscellaneous and mathematical geology; plate tectonics.

INTRODUCTION

Two important goals of my sophomore-level computer-applications course are to develop students' abilities to solve problems and to provide them with the computational skills to facilitate problem solving. Students commonly struggle to convert word problems into mathematical expressions and to draw conclusions from data. To help students develop these skills, I teach a unit on regression analysis that focuses on the scientific process of formulating a question or hypothesis, identifying or gathering data to quantify the problem, analyzing the data using regression techniques, then using the results to make predictions and test the hypothesis.

Many sets of scientific data are well described by linear relationships. The power of regression analysis is that students can visualize the data and relationships in graphs while learning to explain those relationships with mathematics. To reduce the amount of new information students must learn, it is important when teaching regression techniques to choose a problem in which the scientific concepts are well understood by the students. Nearly every student in introductory geology learns how movement of the Pacific plate over the Hawaiian hotspot formed the linear Hawaiian-Emperor volcanic chains (Figure 1). This basic familiarity with the development of the volcanic chain simplifies teaching about linear regression and motivates students to learn more. In this unit, students work with data gathered from the literature to recreate the fundamental work that documented the linear age progression of the Hawaiian-Emperor island chain and provided key information on plate motions.

FITTING A LINE TO DATA

To introduce the concept of regression in my course, I begin with a review of basic algebraic operations and the technique for determining the equation of a line from two points. In a linear relationship, the dependent variable, y, is related to the independent variable, x, by a simple equation:

$$y = ax + b \quad , \qquad (1)$$

where a is the slope and b is the y-intercept of the line. Working in groups during class, students solve problems that have only two points. I use the experience to reinforce the meaning of slope, to emphasize the units assigned to each variable, and to demonstrate the predictive capabilities of a function. Students learn to use equation (1) and known x values to predict new values of y. In addition, they review techniques for finding the intersection of two lines. The problems are simple enough that they can be analyzed with paper and pencil both graphically and analytically.

The main objective of many scientific investigations is to establish relationships, in the form of mathematical equations, that make it possible to predict one or more variables from other known variables. Ideally, we want to predict one quantity exactly in terms of another (as in the two-point problem), but that is seldom possible when multiple data points are involved. More commonly, we predict the average value or expected value. Predicting the average value of one variable in terms of another variable is best done using regression analysis. With regression, one can determine the best-fit line that minimizes the separation between known or observed values, yo_i, and the predicted values, y_i, determined with the line $y_i=ax_i+b$. We can minimize the function:

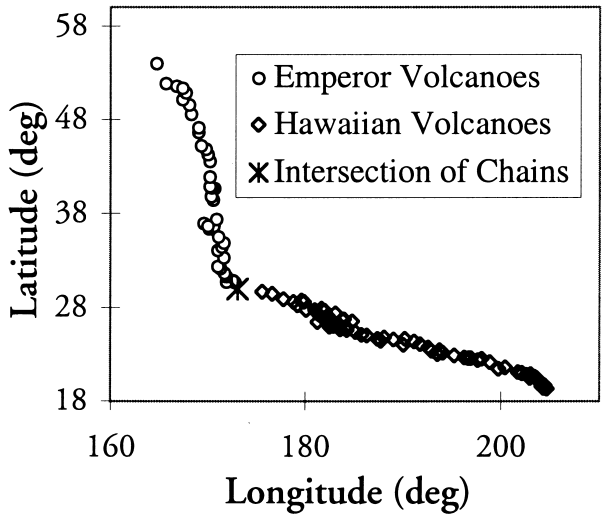

Figure 1. The linear trend of the Hawaiian and Emperor volcanic chains and the calculated point of intersection.

$$\sum_{i=1}^{n}(y_i - yo_i)^2 \quad , \qquad (2)$$

by finding the partial derivatives of (2) with respect to a and b (Freund, 1979).

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n}(y_i - yo_i)^2\right)=0 \quad , \qquad (3)$$

and

$$\frac{\partial}{\partial b}\left(\sum_{i=1}^{n}(y_i - yo_i)^2\right)=0 \quad . \qquad (4)$$

Expanding equation (3), produces

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n}(y_i^2 - 2y_i yo_i + yo_i^2)\right)=0 \quad , \qquad (5)$$

and

$$\sum_{i=1}^{n}\left(2y_i\frac{\partial y_i}{\partial_a} - 2yo_i\frac{\partial y_i}{\partial_a}\right)=0 \quad . \qquad (6)$$

Substituting ax_i+b for y_i , gives

$$2\sum_{i=1}^{n}\left((ax_i+b)x_i - yo_i x_i\right)=0 \quad . \qquad (7)$$

Rearranging terms results in

$$\sum_{i=1}^{n} yo_i x_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i \quad , \qquad (8)$$

which is an algebraic equation that can be solved using a simple calculator. Following the same procedure for equation (4) produces a second algebraic equation

$$\sum_{i=1}^{n} yo_i = a\sum_{i=1}^{n} x_i + nb \quad . \qquad (9)$$

Using equations (8) and (9), students solve several problems in small groups during class and for homework. Example problems include determining relationships between tree age and circumference, lithostatic stress and depth, or elevation and precipitation.

Armed with an understanding of the mathematical basis of regression, students can use one of the many software applications available to perform more challenging regression analyses. These tools are simple and powerful but are only a "black box" approach to problem solving if used with little understanding of the mathematics behind them. Students use Microsoft *Excel* and the function LINEST to determine the values of a and b for a linear function given an array of known x and y values. LINEST can also be used to calculate polynomial functions up to order six. The LINEST function provides statistics for error analysis and evaluation of the quality of fit (Freund 1979; Orvis, 1996), including the standard error of the coefficients a and b, the standard error of the y estimate, F, the sum of the regression, the sum of the residuals and the correlation coefficient, r squared.

DETERMINING PLATE MOTIONS AND AGE PROGRESSION IN THE HAWAIIAN EMPEROR CHAIN

Data Available

To determine a regression line and analyze the quality of the solution, it is essential to understand the data, its variability and sources of errors. For this analysis, we use data on volcano location, age, and distance from the hotspot that were compiled from work of many others by Shaw and others (1980). The latitude and longitude of the volcanoes was determined from seafloor bathymetry by Chase and others (1973), and the given location represents the core of the volcanic edifice. Radiometric age data are only available for 33 of the 104 volcanoes. The age of the volcanoes does not always increase with distance from the hotspot for several reasons. First, rocks sampled on one volcano may have been from a flow that originated on another volcano. Such a relationship occurs today with flows from Mauna Loa encroaching on Kilauea. Second, differential erosion of the volcanoes has exposed different levels of the volcano. At one site, the rocks sampled may have been deposited late in the life of the volcano, while at another site, they may have been deposited very early. Third, hotspot activity can produce flows on more than one volcano at a time. For example, on the island of Hawaii, the volcanoes Mauna Loa, Hualalai, Kilauea, and Loihi are all still active. Fourth, ages may be inaccurate due to poor sample quality or analysis errors.

Shaw and others (1980) report distances from the hotspot as measured along the trends of the Hawaiian and Emperor chains. Distances along the Hawaiian chain are measured from Kilauea to each volcano up to the bend in the chain. The distances to the volcanoes in the Emperor chain are the total distance between Kilauea and Kanmu along the Hawaiian trend plus the distance from Kanmu to the selected volcano along the trend of the Emperor chain. This

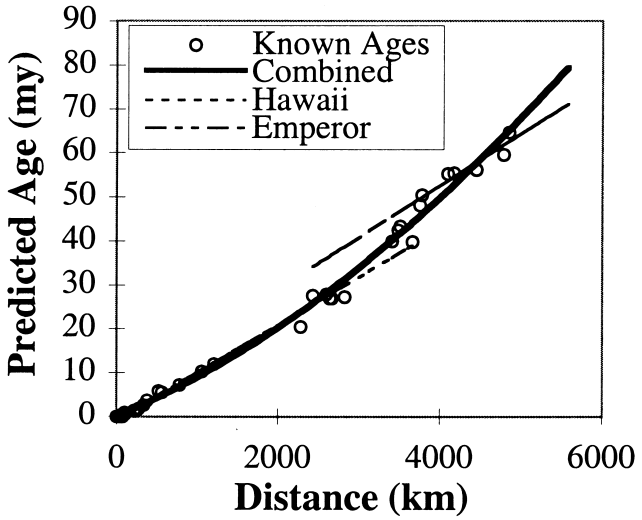

Figure 2. Students use the relationship between Distance from the hotspot and Age of the volcanoes to determine the rate of plate motion during the formation of the Hawaiian and Emperor chains using linear regression. Combining the data for both chains, they also fit a second-order polynomial to the data and compare the results.

has the net effect of projecting the distances versus age onto a single vector (Figure 2). The distance of individual volcanoes from the hotspot increases somewhat episodically rather than linearly with age. Volcanoes appear to have developed along short arcuate segments that are sub-parallel to the general trends of the Hawaiian-Emperor chain. These natural variations in the age and distance from the hotspot data make it difficult to develop a relation between these variables without using a regression technique.

The Problem

There are over 100 hotspots scattered around the globe on both continental and oceanic lithosphere. Their origin is an enigma not well explained by plate tectonic theory. Wilson's (1963) idea as to the mechanism for development and maintenance of a hotspot still sparks debate today (Helmberger and others, 1998; Russel and others, 1998). Geoscientists recognized early in the plate tectonic revolution that about 18 of the hotspot locations are relatively stable and can be used to determine plate motions (Atwater and Molnar, 1974). The linear trace and accessibility of the Hawaiian-Emperor chain has made it a focus of studies of plate motions (Figure 1).

In this activity, students are challenged to: (1) determine the rate and direction of plate motion during the development of the entire Hawaiian-Emperor chain and each segment using both linear and polynomial regression techniques; (2) calculate the location (latitude and longitude) and age of the bend in the island chain; (3) predict the location of Kilauea volcano 10 My from now; and (4) evaluate the quality of the results obtained in each step of the activity. The activity, data, and instruction for using the LINEST function in *Excel* are available at http://www.geo.arizona.edu/K-12/regression/.

Plate motion rate during the Hawaiian chain formation	9.25 ± 2.1 cm/yr
Plate motion rate during the Emperor chain formation	8.53 ± 1.3 cm/yr
Location of the bend	173.05 W 29.92 N
Age of the bend	38.09 ± 1.54 my

Table 1: Results of analysis.

The exercise requires students to have a clear understanding of dependent and independent variables, slope, and the statistical analysis provided as part of the regression calculation. The activity guides students through three steps in performing a regression. In Part I, the students plot and visually inspect the data to estimate the type of function (linear, polynomial, or exponential) needed to explain the data. Then, using a spreadsheet or statistics software, they calculate the coefficients of the particular equation that best fits the data. To determine the rate and direction of plate motion during the development of the Hawaiian-Emperor chain, students model the relationship between age and distance for each segment using a linear regression. They use the results to determine the rate of plate motion through time (Figure 2). In Part II, they model the combined data from both segments with a second-order polynomial with good results. To complete the process, they examine the statistics provided with the regression analyses and evaluate the goodness of the fit of the different functions.

A comparison of regression results indicates that the Hawaiian chain is well fit by a linear function. The smaller number of dates and the more complex age distribution in the Emperor chain result in a less well constrained but good fit to the data. The analysis shows that there was a difference in rates of plate motion during formation of the Hawaiian and Emperor chains (Table 1) and that the plate motion may have increased over time. The overlap of the errors for the two rates suggests that the plate could have been moving faster than 85 mm/yr during the creation of the Emperor chain, but it could not have been going as slow as 85 mm/yr during formation of the Hawaiian chain. The second-order polynomial provides an excellent fit to the combined data based on the correlation coefficient and standard errors of the coefficients. Students use the new age relationships to calculate predicted ages for all the volcanoes in the island chain. (Note that the rate of plate motion determined with the polynomial $y=cx^2+ax+b$ is not equal to $1/a$, as in the linear regressions, but is a function of both a and c.) A comparison of the predicted ages and the measured ages presents an opportunity to investigate the concept of formal errors and data errors.

Calculating the location of the bend in the island chain in Part III requires calculation of the best-fit line for the latitude and longitude of the volcanoes in each segment of the chain and simultaneous solution of the two equations to find the point of intersection. In Part IV, students use their algebraic and computational skills with their understanding of regression

to develop a relationship between age and location of volcanoes. To determine a relationship between the age and location (latitude and longitude) of the volcanoes requires the students to first determine the relationship between age and longitude, then to determine the relationship between longitude and latitude. The equations have the following form:

$$y(Longitude)=a*x(Age)+b \quad , \qquad (10)$$

$$y(Latitude)=a*(Longitude)+b \quad . \qquad (11)$$

Solving equation (10) and substituting the results into equation (11) provides the ability to predict the location of a particular volcano given a known time in the past or future. In addition, by substituting the longitude of the bend in the chains into the results from equation (10) and rearranging terms, the student can determine the age of the bend. When finished, students will have applied many of the basic mathematics skills to solving an important geologic problem. Hopefully, they will also have a greater appreciation for the importance of mathematics in science.

ASSESSMENT

I have continually evaluated and improved this unit over the past five years based on student comments, classroom observations, and analysis of student answers. The overall structure and problems presented in the activity have not changed significantly; however, the clarity of the instructions and questions has improved dramatically. The success of the activity is highly dependent on the classroom instruction that builds student understanding of regression and its uses. The biggest challenge for students is in converting the word problem into a mathematical expression. They also face difficulties in using the regression results to make predictions about the system they are modeling. Students also have difficulty understanding the difference between the dependent and independent variables.

It is essential to review the basic algebraic and geometric relationships used in this exercise with small data sets and in-class problem solving. Most students have encountered and mastered these techniques in previous math courses but never actually applied them to a real problem. The classroom examples must also emphasize the predictive value of a regression equation. A recent survey of student comments on this unit is quite favorable. While many students felt the activity was challenging, they also indicated that they gained a clear understanding of regression analysis and its usefulness for modeling data. Further, they enjoyed the process of recreating a piece of familiar and interesting scientific research.

CONCLUSIONS

The linear regression technique provides excellent results for determining the age progression and location of the Hawaiian-Emperor volcanic chain. Using a geologic process that is well understood by even the beginning student, we are able to introduce students to problem solving using real data and standard analysis techniques. Students typically begin this unit with little understanding of regression or its application to problem solving. However, over the course of this activity, students build on their basic understanding of the linear equations to develop a deeper understanding of regression as a tool for data analysis and testing models. The struggle with these data and problems helps them realize how the mathematics is useful in their work.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation grant number EHR-9451649. The author appreciates all the students who have provided feedback on this activity, especially Marie Renwald and Anne Paquette. The activity and data used for this exercise are available at http://www.geo.arizona.edu/K-12/regression/. An answer key is available by contacting the author directly.

REFERENCES

Atwater, T. and Molnar, P., 1974, Relative motion of hot spots in the mantle: University of California, Scripps Institute of Oceanography Contributions, v. 44, part 2, p. 1632-1635.

Chase, T.E., Menard, H.W., and Mammerickx, J., 1973, Bathymetry of the north Pacific, charts 1-10: La Jolla, California, University of California, Scripps Institute of Oceanography.

Freund, J.E., Modern elementary statistics (5th edition): Englewood Cliffs, New Jersey, Prentice Hall, 510 p.

Helmberger, D.V., Wen, L., and Ding, X., 1998, Seismic evidence that the source of the Iceland hotspot lies at the core-mantle boundary: Nature, v. 396, p. 251-255.

Orvis, W.J., Excel for scientists and engineers, 2nd edition: San Francisco, California, Sybex, 547 p.

Russel, S., Lay, T., and Garnero, E.J., 1998, Seismic evidence for small-scale dynamics in the lowermost mantle at the root of the Hawaiian hotspot: Nature, v. 369, p. 255-257.

Shaw, H., Jackson, E.D., and Bargar, K.E., 1980, Volcanic periodicity along the Hawaiian-Emperor Chain, American Journal of Science, v. 280-A, p. 667-708.

Wilson, J.T., 1963, A possible origin of the Hawaiian Islands: Canadian Journal of Physics, v. 41, p. 863-870.

ABOUT THE AUTHOR

Michelle Hall-Wallace enjoys using mathematics and technology to investigate geologic problems. She appreciates all the hard work of the geologists who gather the critical data, but she will trade her rock hammer for a computer any day.