GS543 Tutorial-4

Bivariate analysis aims to understand the relationship between two variables,

x and y. Examples are the length and width of a fossil, the sodium and potassium content of volcanic glass, and the organic matter content along a sediment core. When the two variables are measured on the same object, x is usually identified as the independent variable and y as the dependent variable. If both variables have been generated in an experiment, the variable manipulated by the experimenter is described as the independent variable.

In some cases, neither variable is manipulated and neither is independent.

Agedepth text file is example of a bivariate data set. The thirty data points represent the *age* of a sediment (in kiloyears before present) at a certain *depth* (in meters) below the sediment-water interface. The combined distribution of the two variables suggests a linear relationship between *age* and *depth*, i.e., the rate of increase in the sediment age with depth is constant. Linear regression yields the equation *age*=b_o+ b_1 *depth*, indicating an

increase in sediment age of m kyrs per meter of sediment depth (the slope of the regression line).

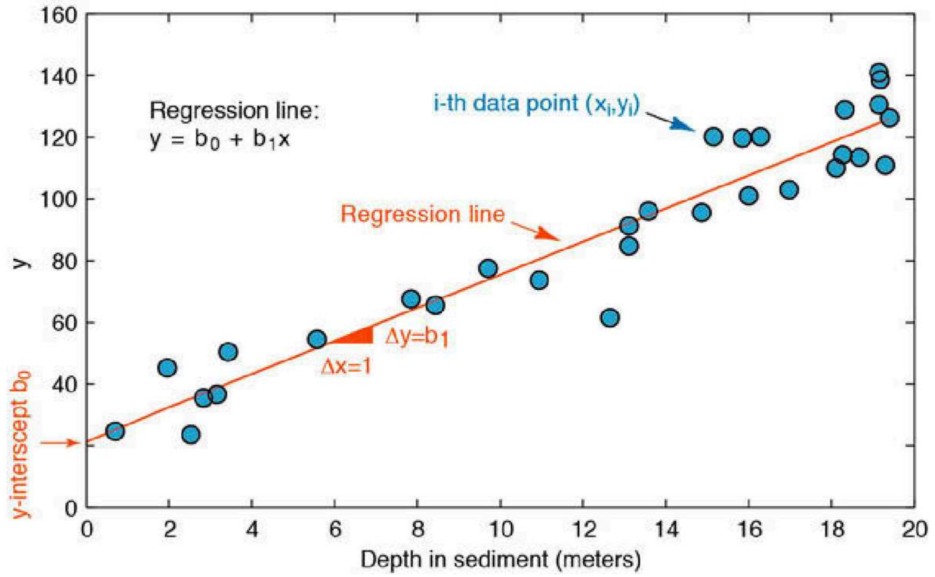
Write down a computer program to determine b_o & b_1.

Understand the Problem : Classical Linear Regression Analysis

Linear regression offers another way of describing the relationship between

the two variables x and y. Classical regression assumes that y responds to x and that the entire

dispersion in the data set is contained within the y-value (as shown in the image). This means that x is then the independent variable (also known as the predictor variable, or the regressor). The values of x are defined by the experimenter



and are often regarded as being free of errors. An example is the location x within a sediment core from which the variable y has been measured. The dependent variable y contains errors as its magnitude cannot be determined accurately. Linear regression minimizes the deviations Δy between the data points xy and the value y predicted by the best-fit line $y=b_0+b_1x$ using a least-squares criterion. The basic equation for a general linear model is

$$y = b_0 + b_1x$$

where b_0 and b_1 are the regression coefficients. The value of b_0 is the intercept with the y-axis and b_1 is the slope of the line. The squared sum of the Δy deviations to be minimized is

$$\sum_{i=1}^{n}\left(\Delta y_i\right)^2 = \sum_{i=1}^{n}\left(y_i - (b_0 + b_1x_i)\right)^2$$

Partial differentiation of the right-hand term in the equation and setting it to zero yields a simple equation for the regression coefficient b_1:

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

The regression line passes through the data centroid defined by the sample means, and we can therefore compute the other regression coefficient b_0,

$$b_0 = \bar{y} - b_1\bar{x}$$

using the univariate sample means and the slope b_1 computed earlier.