Csci 1523
Spring 2016
Developing a Linear Regression Model
Due April 3, 2016 @ 11:59p

This programming project contains 6 pages (including this cover page) and includes 2 programming problems.

Before starting to work on this laboratory please read through each programming problem carefully making sure that all of the examples and figures are present and that you have the correct number of pages in the assignment.

In our course we encourage collaboration on programming problems however exchanging SOURCE CODE is expressly prohibited.

To enable your collaboration we have established on the course site a discussion forum for this particular assignment. students are encouraged to post questions concerning the documentation, logic, and coding of the assignment on this site. however as mentioned sharing source code is strictly prohibited.

Important: File submission standards for Dropbox

Please note that all Python files submitted to the Dropbox for this assignment should should have a .txt file extension. Any programs submitted with a file extension other than .txt will not be graded.

These files should be internally documented as we have stipulated for programs in our course.

As is the policy in our course if a file is submitted with an extension other than .txt it may be submitted after the due date with a 50% reduction in credit.

Programming Project 1: Developing a linear regression of data

1. Developing the slope and intercept of a linear regression equation based on sample data

   In this programming assignment we will develop a simple linear regression model to predict the values of one variable based on the values of a second variable. Variable we are predicting is called the *criterion variable* and is referred to as the y, the variable we base our predictions on is called the *predictor variable* and is referred to as x. When there is only one predictor variable the prediction method is called simple regression. Our programming assignment is to develop a simple regression between two variables x and y.

   These regression equations are typically developed using specialized statistical software. In our case we will develop the specialized statistical software to do the regression.

   In Figure 1 we have a plot of raw data. This data is typically taken from a data file which contains sample x and y data.

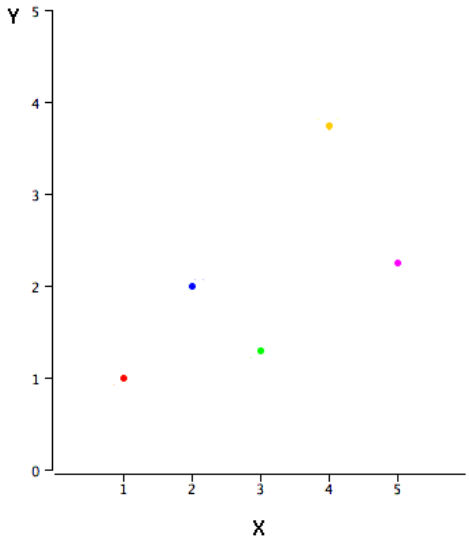

Figure 1: Sample data plotted in 2 dimensions

In Figure 2 we show the raw data and a regression line plotted.

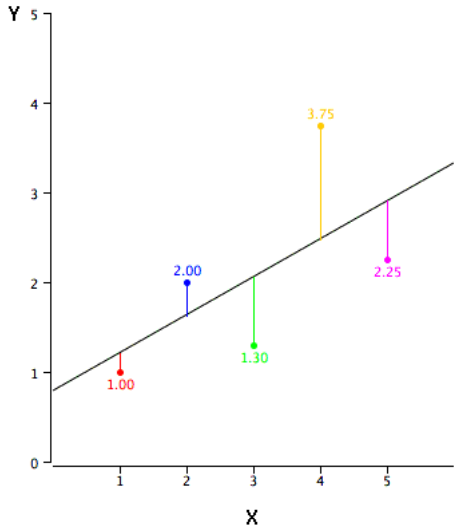

Figure 2: Sample data with regression line

The purpose of this programming assignment is to:

1. Read raw data from a text file. The file **Regression.dat** can be found in the Dropbox for this programming assignment.
2. **Required:** Using the raw data and the mathematics necessary calculate the simple regression line:
   (a) Determine the slope of the line, m.
   (b) The y-intercept, b.
   (c) The value of the coefficient of determination, R^2.
3. **Bonus Part:** Plot the raw data and the regression lines showing the error bars as shown in Figure 2.

Calculating the slope and intercept of the regression line.

The regression line should represent the "best fitting" line based on the data set. By "best fitting" we mean the line through the sample data which minimizes the sum of the distances from the regression line to the sample points. This involves calculating the distance from each sample point to the regression line totaling them and ensuring that this total represents the minimum summation of these distances.

Fortunately for us this problem has been studied at great length and we can simply use the equations for calculating the regression line based on those provided

to us from the field of statistics.

The regression line has the form:

$$y = mx + b \tag{1}$$

where:

- y is the value we are predicting
- x is the predictor variable
- m is the slope of the regression line
- b is the intercept point of the regression line on the Y axis

Our programming problem is to use the data set provided to estimate the values for m and b.

Statistics provides us the following equations which allow for the direct calculation of these variables.

First the slope, m, is calculated using the following equation:

$$m = \sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})] / \sum_{i=1}^{n}[(x_i - \bar{x})^2] \tag{2}$$

where:

- m is the slope of the regression line
- x_i is the value of x from the dataset
- y_i is the value of y from the dataset
- $\bar{x}$ is the average value of x data
- $\bar{y}$ is the average value of y data
- n is the number of data points in the dataset

From Equation 2 it is very easy to calculate b using the following equation:

$$b = \bar{y} - m\bar{x} \tag{3}$$

The variables are as defined in equation 1 and equation 2 given above.

2. Calculating the coefficient of determination, R^2

Statistical predictors are of very little use to us unless we have some type of an idea concerning how well our regression equation can predict the *criterion variable*, y.

The coefficient of determination, R^2, is a statistical means of determining how well our regression equation fits the data provided to us.

This is a relatively straightforward calculation which is given in equation for below:

$$R^2 = \{(1/n) * \sum_{i=1}^{n}[(x_i - \bar{x}) * (y_i - \bar{y})]/(\sigma_x * \sigma_y)\}^2 \tag{4}$$

Where:

x_i is the value of x from the dataset

y_i is the value of y from the dataset

$\bar{x}$ is the average value of x data

$\bar{y}$ is the average value of y data

n is the number of data points in the dataset

σ_x is the standard deviation of the x values

σ_y is the standard deviation of the y values

The standard deviation of the x and y values is a relatively straightforward calculation as well. These are given below, using the variables as defined above:

$$\sigma_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/n} \tag{5}$$

$$\sigma_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2/n} \tag{6}$$

Students should calculate R^2 for the dataset provided and the value should be output to the screen.

Suggested approach to the problem: Students are free to develop the program which calculates these values as they see fit. However the following initial steps may aid in your program development:

1. As we have done in our classroom activities and shown in the videos, open the data file and extract the data into two different lists. One list should contain the y values represented as float types the other list should contain the x values also represented as float types.
2. Calculate the average values for x and y based on the contents of each list and print these to the screen to ensure you are calculating them correctly.

3. Using the list which contains the sample data in float values, utilize loops to loop through the lists and perform the calculations as shown in equation.
4. Output the values of m, b and R^2 to the terminal window.
5. **Bonus Part:** Using the data and the equations for the regression developed above use Zelle's graphic package to develop a plot of the data, the regression lines and error bars. As mentioned Figure 2 is an example of what your final graphic should look like.

SUGGESTION: We all get a bit rusty with our fundamental mathematics and I have found a webpage which contains a video which should help you in refreshing your understanding. It can be accessed at:

`http://stattrek.com/regression/regression-example.aspx`

I feel the accompanying video is excellent even though it is developed for high school students. It is very thorough and I believe will be a help to some of you.

I hope this is of some assistance to you in working through the assignment.

SUBMISSION FILE NAMES:

1. Submit the program for the assignment in a file called:

   `Csci1523Assign2.txt`

2. Submit the bonus part of the assignment in a file called:

   `Csci1523Assign2Regression.txt`