CS 8, Winter 2015
Homework Assignment #? (draft)

Assignment Overview

This assignment will give you more experience on the use of functions and dictionaries. You will practice them by processing a file from a real-life dataset. Given a data file of 507 individuals and their physical attributes (weight, height, etc. from the body dataset), create two linear regression models (explained below):

- Between a person's BMI and their age.
- Between a person's weight and a combination of physical attributes. It is proposed that the following formula be used for the physical attributes:
    - -110 + 1.34(ChestDiameter) + 1.54(ChestDepth) + 1.20(BitrochantericDiameter) + 1.11(WristGirth) + 1.15(AnkleGirth) + 0.177(Height)

Background

BMI, short for Body Mass Index, is a measure based on a person's weight and height. It is used as an estimator of healthy body weight (see http://en.wikipedia.org/wiki/Body_mass_index).

Linear regression is a form of statistical analysis in which the relationship between one or more independent variables and another variable, called the dependent variable, is modeled by a least squares function, called a linear regression equation. A linear regression equation with one independent variable represents a straight line when the predicted value (i.e. the dependant variable from the regression equation) is plotted against the independent variable. For example, suppose that a straight line is to be fit to the points (y_i, x_i), where $i = 1, \ldots, n$; y is called the **dependent variable** and x is called the **independent variable**, and we want to predict y from x.

Least Squares and Correlation

The method we are going to use is called the least squares method. It takes a list of x values and y values (the same number of each) and calculates the slope and intercept of a line that best matches those values. See http://easycalculation.com/statistics/learn-regression.php for an example.

To calculate the least squares line, we need to calculate the following values from the data:

- `sumX` and `sumY`: the sum of all the X values and the sum of all the Y values
- `sumXY`: the sum of all the products of each corresponding X,Y pair
- `sumXSquared` and `sumYSquared`: the sum of the square of every X value and the sum of the square of every Y value
- `N`: the number of such (x-y) pairs

The calculation then is:

- `slope = (N*sumXY - (sumX*sumY))/(N*sumXSquared - (sumX)²)`
- `intercept = (sumY – (slope*sumX)) / N`

We will also calculate the correlation coefficient, and indication of how "linear" the points are (how much, in total, the points are correlated as a line). That calculation is:

- ```
  corr = (N*sumXY - (sumX*sumY)) /
              sqrt((N*sumXSquared - (sumX)²) * (N*sumYSquared - (sumY)²))
  ```
  ```

The correlation value ranges between -1 and 1. A negative value means an inverse correlation, a positive value a positive correlation. Values near -1 or 1 are "good" correlations, values near 0 are "bad" correlations. See http://easycalculation.com/statistics/learn-correlation.php

Assignment Description

- The data come from file `'body.dat'`. The file `'body.txt'` describes the data.
- For the BMI calculation, age will be the `x` values, and BMI the `y` values.
    - BMI is not a value found in the data. You will have to calculate it using the data.
    - Get the units right when you calculate the BMI!
- For the formula calculation, the physical attribute formula will be the x values and weight will be the y values.
    - All units are correct as provided in the data for the formula
- Plot all data points (in black) as a scatter plot. Calculate the slope and intercept of a linear regression line for those two measures (age against BMI and physical attribute formula result against weight), and overlay the line (in red) on the scatter plot. Output the correlation coefficients on the console.

.

Assignment Deliverables

The deliverable for this assignment is the following file:

body.py – the source code for your Python program

Be sure to use the specified file name and submit it for grading via the **turnin** system before the project deadline.

Assignment Notes:

1. Your program should comprise the following functionalities: open and parse input files, compute BMI and the particular physical attribute according to the given formula, draw scatter plot, compute regression lines, and plot regression lines. All these functionalities should be invoked appropriately by one function called "regression" which accepts one single argument that is the name of the input file.

2. Don't try to tackle this project all at once. Complete one function (or part of a function) and test it out.

3. Test your least squares function on known data to make sure it works.

4. You should ***test your functions*** before using them in the program. Create some small lists of known x and y values, for example [1,2,3,4,5] for both x and y. The slope and intercept of that should be obvious, as should the correlation. If you don't get the required answers, fix the function before moving on. Create a small data file with only two or three entries and test that you can parse it correctly. Testing functions will make your life easier.

Sample Output:

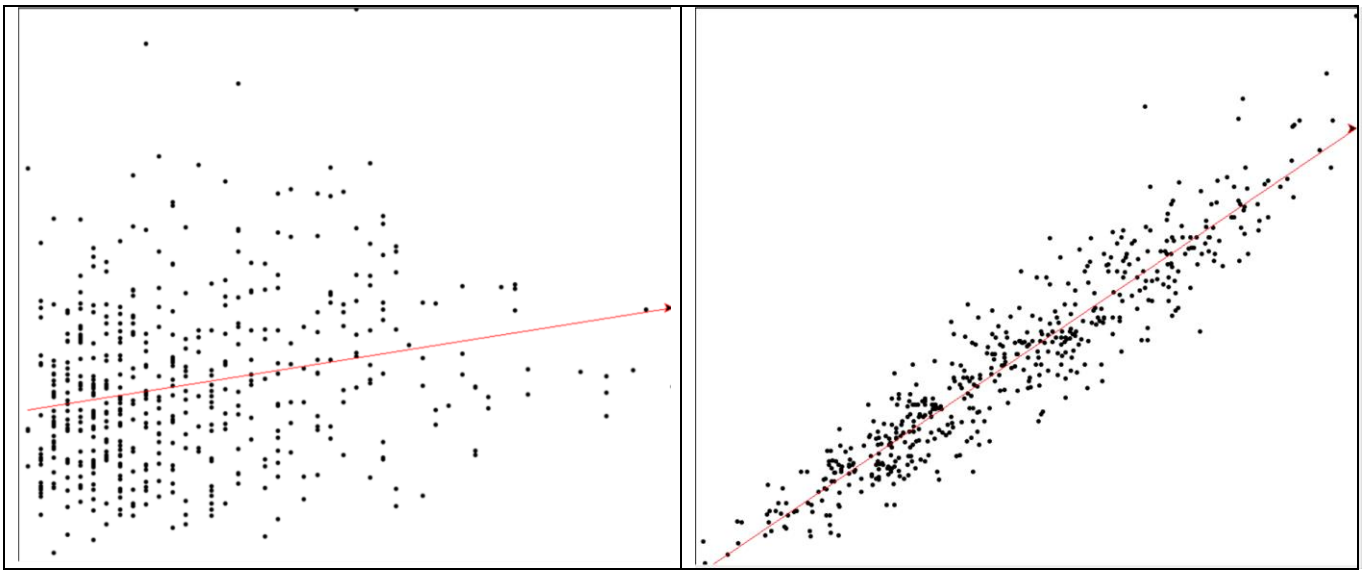

Figure 1: left: age vs. BMI scatter plot and regression line, right: physical attribute vs. weight scatter plot and regression line.