Advanced Programming (FEB23007-14)

1. Exercise

Deadline for submission: 2015-01-25 23:59 CET

Instructions

Include in each source file (in class documentation, @author) your names and student numbers. Submit the exercise as a zip file containing _only_ the source files in root of the zip. Submit via blackboard. Note that incorrectly submitted or non-compiling exercises are automatically awarded 0 points. Remember to document your code with Javadoc-annotations.

Exercise

Let us consider a data set of n 2-dimensional points $\{x_i, y_i\}$, with $i \in \{1, \ldots, n\}$. Suppose we would like to find a linear equation $y = a + bx$ best describing this data set. Here, the best line minimizes the sum of its squared residuals with respect to the original data. When creating a simple linear regression model by using an ordinary least squares estimator, estimations of a and b, i.e., $\widehat{a}$ and $\widehat{b}$, respectively, can be computed as

$$\begin{aligned}\widehat{b} &= \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2},\\ \widehat{a} &= \overline{y} - \widehat{b}\overline{x},\end{aligned}$$

with $\overline{x}$ and $\overline{y}$ representing the sample averages of x and y, respectively. Download the data set from `http://smaa.fi/static/prog3/2012/data/food.dat`. The data set consists of observations of household weekly incomes (in 100 $'s, first column) and food expenditures (in $'s, second column) and other two variables. Write a Java program that reads in the data set, computes estimations $\widehat{a}$ and $\widehat{b}$, and displays $\widehat{a}$, $\widehat{b}$, and the Sum of Squared Residuals (SSR).

For making the program in an object-oriented manner, you should implement (at least) the following classes:

- DataSet: a class for modeling a data set. The class should have a constructor that takes a file name containing the data set as an input. Do not handle possible exceptions within the DataSet, but throw them to be handled by the class's user. In addition to the file not existing or not being readable, you might have a data file with unequal amount of columns in the different rows. Note that the class should read in all data, not just the part used for the regression. It should have a method for obtaining a column of data.
- LinearRegression: a class for computing the regression model with the given two vectors of doubles representing the dependent and independent variable observations. Include accessor methods for obtaining the regression coefficients and the SSR.
- A program entry point class with the name Main that performs simple linear regression on the food.dat data set, using the second column as the dependent variable (y) and the first column as the independent variable (x). This class should use DataSet and LinearRegression for loading the data and computing the regression coefficients, respectively.

Hints

- Remember to clearly separate the classes' responsibilities. E.g. DataSet only reads in and stores the data set in a rectangular format where each column is a single variable.
- Remember to clearly distinguish accessor- and mutator-methods.
- java.util.StringTokenizer allows easy split of a single string to its components, and java.io.StreamTokenizer similar but more complex functionality with Readers (allows to skip comments as well).