Voortgezet Programmeren (FEB23007-11)

2. Exercise

Deadline for submission: 2012-01-22 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 $y = ax + b$ 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

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

with $\overline{x}$ and $\overline{y}$ representing the sample averages of x and y, respectively. Download the data set from `http://smaa.fi/tommi/courses/prog3/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 the data set, computes estimations $\widehat{a}$ and $\widehat{b}$, and displays $\widehat{a}$, $\widehat{b}$, and the sum of squared residuals.

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 a data set as an input. Handle in an appropriate manner (=catching it + displaying an error message) possible exceptions (e.g. missing file or incorrect file contents' format). Note that the class should read in all data, not just the part used in the regression.

- LinearRegression: a class for computing the regression model with a given DataSet and column indices of dependent and independent variables. The constructor should take at least a DataSet as a parameter.

- A tester class with name Main that performs simple linear regression on the food.dat data set, while 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, respectively.

Hints

- Remember to clearly separate the classes' responsibilities. E.g. DataSet only reads and represents the data set in a matrix-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).