Program 4

CS 140
Spring 1994
Craig A. Rich

Scientists and engineers frequently perform experiments designed to provide measurements of two variables x and y (e.g., I.Q. and test scores). They often compute measures of central tendency (such as the mean) and measures of dispersion (such as standard deviation) for each variable, and then attempt to see whether or not there is any relationship between the variables. If there is a relationship between x and y that is describable using a linear equation of the form $y = mx + b$, the data is said to fit a linear curve.

For this program, the data to be input is in the file `[cs.carich]program_4.dat` in my directory. This file contains more than one group of data, so the statistical computations described below will be performed more than once (e.g., in a `for`-loop). The data to be input will be of the following general form, and your procedure should work correctly on any data file of that form, not just the one provided in my directory.

g	
n_1	
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_{n_1}	y_{n_1}
n_2	
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_{n_2}	y_{n_2}
$\vdots$	
n_g	
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_{n_g}	y_{n_g}

g is the number of groups of data. n_i is the number of pairs of data in group i. (x_1, y_1), (x_2, y_2), ..., (x_{n_i}, y_{n_i}) are the pairs of data in group i.

The required computations for a group of data can be performed as follows:

$$\begin{aligned}
\texttt{Sum_X} &= \sum_{j=1}^{n} x_j; \\
\texttt{Sum_Y} &= \sum_{j=1}^{n} y_j; \\
\texttt{Sum_XY} &= \sum_{j=1}^{n} x_j y_j; \\
\texttt{Sum_X_Squared} &= \sum_{j=1}^{n} x_j^2; \\
\texttt{Sum_Y_Squared} &= \sum_{j=1}^{n} y_j^2; \\
\texttt{Mean_X} &= \texttt{Sum_X}/n; \\
\texttt{Mean_Y} &= \texttt{Sum_Y}/n; \\
\texttt{Standard_Deviation_X} &= \sqrt{\texttt{Sum_X_Squared}/n - \texttt{Mean_X}^2}; \\
\texttt{Standard_Deviation_Y} &= \sqrt{\texttt{Sum_Y_Squared}/n - \texttt{Mean_Y}^2}.
\end{aligned}$$

The slope m and y-intercept b of the best-fit line $y = mx + b$ can be computed using the "least squares" method:

$$\begin{aligned}
m &= \frac{\texttt{Sum_XY} - n * \texttt{Mean_X} * \texttt{Mean_Y}}{\texttt{Sum_X_Squared} - n * \texttt{Mean_X}^2}; \\
b &= \texttt{Mean_Y} - m * \texttt{Mean_X}.
\end{aligned}$$

The output of this procedure should be written to a file named `program_4.out`, and should give the information about each group of data formatted *exactly* as follows:

```
Statistical measures for group 1

Mean X:                      9.57
Mean Y:                      8.86
Standard deviation X:        0.73
Standard deviation Y:        1.17

Best linear fit: Y = 0.78 * X + 1.37


Statistical measures for group 2

Mean X:                    133.12
Mean Y:                    219.85
Standard deviation X:       20.12
Standard deviation Y:       27.37

Best linear fit: Y = 1.22 * X + 58.02
```

Hand in one contiguous printout containing a compiler listing of `program_4.ada` and a listing of the output obtained on input file `[cs.carich]program_4.dat`.