Requerimientos del Programa 3

Utilizando el **PSP2**, escribe un programa que:

- Lea del teclado el nombre de un archivo
- Lea de este archivo lo siguiente:
  - El primer renglón contiene un número real mayor o igual a cero, al cual llamaremos x_k.
  - A partir del segundo renglón habrá en cada renglón una pareja (x, y) de dos números reales mayores o iguales a cero, separados por una coma
  - El fin del archivo marca el final de las parejas de datos
- Calcule los siguientes datos
  - La cantidad de parejas de datos leídas (N)
  - Los coeficientes de correlación $r_{x,y}$ y r^2
  - Los parámetros de regresión lineal β_0 y β_1
  - Una predicción mejorada y_k, donde $y_k = \beta_0 + \beta_1 x_k$
- Escriba en pantalla estos valores calculados de acuerdo con el siguiente formato:

```
N  = xx
xk = xx
r  = x.xxxxx
r2 = x.xxxxx
b0 = x.xxxxx
b1 = x.xxxxx
yk = x.xxxxx
```

NOTA:

✓ Los valores de r, r2, b0, b1 y yk se desplegarán con 5 decimales (redondeados hacia arriba en su último dígito, por ejemplo: 0.123455 se desplegará como 0.12346, mientras que 0.123454 se desplegará como 0.12345)

Otras características que ***debe*** cumplir el programa:

- No utilizará ningún GUI para operar (funcionará desde la consola)
- Debe estar construido con programación orientada a objetos
- Debe contar con al menos 3 clases "relevantes" (la clase que contiene el "main" se cuenta como una de estas 3 clases)
- El ***único*** código que puede ser reutilizado es el de tus programas 1 y 2
- Debe manejar apropiadamente (no tronar) ***todas*** las condiciones normales y de excepción
- Debe pasar exitosamente ***todos*** los casos de prueba (*error máximo 0.0001*):
  - Los diseñados por ti en la fase de diseño, y
  - Los siguientes 3 casos de prueba (es obligatorio incluirlos en el Diseño de las Pruebas):

Objetivo de la prueba	Instrucciones y datos de entrada	Resultados Esperados
Probar con una lista de datos	Teclear en pantalla: Arch1.txt	N = 10 xk = 386 r = 0.95450 r2 = 0.91106 b0 = -22.55253 b1 = 1.72793 yk = 644.42938
Probar con una lista de datos	Teclear en pantalla: Arch2.txt	N = 10 xk = 386 r = 0.94803 r2 = 0.89877 b0 = -4.60375 b1 = 0.14016 yk = 49.49938
Probar con una lista de datos	Teclear en pantalla: Arch3.txt	N = 28 xk = 192 r = 0.14775 r2 = 0.02183 b0 = 38.49303 b1 = 0.15615 yk = 68.47322

Fin de los requerimientos

Explicación y ejemplo de cómo se realizan los cálculos (no son requerimientos)

(Tomado del curso original del PSP, del Software Engineering Institute)

Regression

Overview

Linear regression is a way of optimally fitting a line to a set of data. The linear regression line is the line where the distance from all points to that line is minimized. The equation of a line can be written as

$y = \beta_0 + \beta_1 x$

In Figure 1, the best fit regression line has parameters of β_0 = -4.0389 and β_1 = 0.1681.

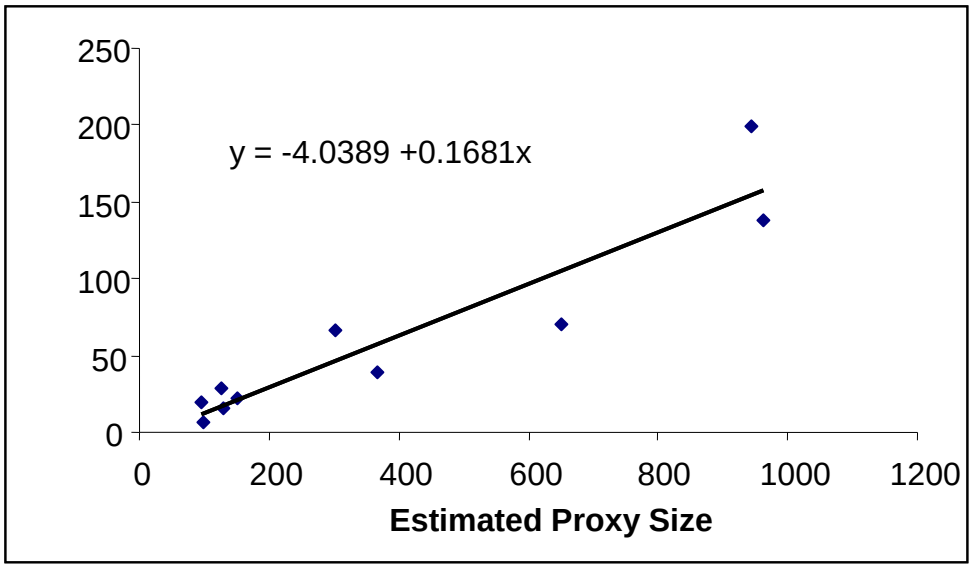

Figure 1

Continued on next page

Using regression in the PSP

Looking at Figure 1, how many hours do you think it would take to develop a program with an estimated proxy size of 500?

Using PROBE method A for time, the estimate would be $TimeEstimate = \beta_0 + \beta_1(500)$ or an estimate of 80.011 hours.

The PSP PROBE method uses regression parameters to make better predictions of size and time based on your historical data.

PROBE methods A and B differ only in the historical data (*x* values) used to calculate the regression parameters. In PROBE method A, **estimated proxy** size are used as the *x* values. In PROBE method B, **plan added and modified** size are used as the *x* values.

PROBE methods for size and time differ only in the historical data (*y* values) used to calculate the regression parameters. To predict improved size estimates, **actual added and modified LOC** are used as the *y* values. To predict time estimates, **actual development times** are used as the *y* values.

Historical Data Used		***x* values**	***y* values**
Size Estimating	PROBE A	Estimated Proxy Size	Actual Added and Modified Size
	PROBE B	Plan Added and Modified Size	Actual Added and Modified Size
Time Estimating	PROBE A	Estimated Proxy Size	Actual Development Time
	PROBE B	Plan Added and Modified Size	Actual Development Time

Correlation

Overview

The correlation calculation determines the relationship between two sets of numerical data.

The correlation $r_{x,y}$ can range from +1 to -1.

- Results near +1 imply a strong positive relationship; when *x* increases, so does *y*.
- Results near -1 imply a strong negative relationship; when *x* increases, *y* decreases.
- Results near 0 imply no relationship.

Using correlation in the PSP

Correlation is used in the PSP to judge the quality of the linear relation in various historical process data that are used for planning. For example, the relationships between estimated proxy size and actual time or plan added and modified size and actual time.

For this purpose, we examine the value of the relation r_{xy} squared, or r^2.

If r^2 is	the relationship is
$.9 \leq r^2$	predictive; use it with high confidence
$.7 \leq r^2 < .9$	strong and can be used for planning
$.5 \leq r^2 < .7$	adequate for planning but use with caution
$r^2 < .5$	not reliable for planning purposes

Limitations of correlation

Correlation doesn't imply cause and effect.

A strong correlation may be coincidental.

> From 1840 to 1960, no U.S. president elected in a year ending in 0 survived his presidency.
> Coincidence or Correlation?

Many coincidental correlations may be found in historical process data.

To use a correlation, you must understand the cause-and-effect relationship in the process.

Calculating regression and correlation

Calculating regression and correlation

The formulas for calculating the regression parameters β_0 and β_1 are

$$\beta_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - \left(n x_{avg} y_{avg}\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \left(n x_{avg}^2\right)}$$

$$\beta_0 = y_{avg} - \beta_1 x_{avg}$$

The formulas for calculating the correlation coefficient $r_{x,y}$ and r^2 are

$$r_{x,y} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\left(\sum_{i=1}^{n} y_i^2\right) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$r^2 = r * r$$

where

- Σ is the symbol for summation
- *i* is an index to the *n* numbers
- *x* and *y* are the two paired sets of data
- *n* is the number of items in each set *x* and *y*
- x_{avg} is the average of the *x* values
- y_{avg} is the average of the *y* values

An example

An example

In this example, we will calculate the regression parameters (β_0 and β_1 values) and correlation coefficients $r_{x,y}$ and r^2 of the data in the Table 3.

n	***x***	***Y***
1	130	186
2	650	699
3	99	132
4	150	272
5	128	291
6	302	331
7	95	199
8	945	1890
9	368	788
10	961	1601

Table 1

$$\beta_1 = \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - \left(n x_{avg} y_{avg}\right)}{\left(\sum_{i=1}^{n} x_i^2\right) - \left(n x_{avg}^2\right)}$$

1. In this example there are 10 items in each dataset and therefore we set *n* = 10.
2. We can now solve the summation items in the formulas.

n	*x*	*y*	x^2	*x*y*	y^2
1	130	186	16900	24180	34596
2	650	699	422500	454350	488601
3	99	132	9801	13068	17424
4	150	272	22500	40800	73984
5	128	291	16384	37248	84681
6	302	331	91204	99962	109561
7	95	199	9025	18905	39601
8	945	1890	893025	1786050	3572100
9	368	788	135424	289984	620944
10	961	1601	923521	1538561	2563201
Total	$\sum_{i=1}^{10} x_i = 3828$	$\sum_{i=1}^{10} y_i = 6389$	$\sum_{i=1}^{10} x_i^2 = 2540284$	$\sum_{i=1}^{10} x_i y_i = 4303108$	$\sum_{i=1}^{10} y_i^2 = 7604693$
	$x_{avg} = \frac{3828}{10} = 382.8$	$y_{avg} = \frac{6389}{10} = 638.9$			

Continued on next page

An example, Continued

An example, cont.

3. We can then substitute the values into the formulas

$$\beta_1 = \frac{(4303108) - (10 * 382.8 * 638.9)}{(2540284) - (10 * 382.8^2)}$$

$$\beta_1 = \frac{1857399}{1074926} = 1.727932$$

$$r_{x,y} = \frac{10(4303108) - (3828)(6389)}{\sqrt{[10(2540284) - (3828)^2][10(7604693) - (6389)^2]}}$$

$$r_{x,y} = \frac{18573988}{\sqrt{[10749256][35227609]}} \qquad r_{x,y} = \frac{18573988}{19459460.1}$$

$$r_{x,y} = 0.9545$$

$$r^2 = 0.9111$$

4. We can then substitute the values in the β_0 formula

$$\beta_0 = y_{avg} - \beta_1 x_{avg}$$

$$\beta_0 = 638.9 - 1.727932 * 382.8 = -22.5525$$

5. We now find y_k from the formula $y_k = \beta_0 + \beta_1 x_k$

$$y_k = -22.5525 + 1.727932 * 386 = 644.4294$$