Requerimientos del Programa 7

Utilizando el **PSP 2.1** escribe un programa que:

- Lea del teclado el nombre de un archivo
- Lea de este archivo lo siguiente:
  - El primer renglón contiene tres números reales mayores o iguales a cero separados con comas, a los cuáles llamaremos w_k, x_k, y_k.
  - A partir del segundo renglón habrá en cada renglón un cuádruplo (w, x, y, z) de números reales mayores o iguales a cero, separados con comas
  - El fin del archivo marca el final de los cuádruplos de datos
- Calcule los siguientes datos
  - La cantidad de cuádruplos leídos (N)
  - Los parámetros de regresión múltiple β_0, β_1, β_2 y β_3
  - Una predicción mejorada z_k, donde $z_k = \beta_0 + \beta_1 w_k + \beta_2 x_k + \beta_3 y_k$
- Escriba en pantalla estos valores calculados de acuerdo al siguiente formato:

```
N  = x
wk = x.xxxxx
xk = x.xxxxx
yk = x.xxxxx
------------
b0 = x.xxxxx
b1 = x.xxxxx
b2 = x.xxxxx
b3 = x.xxxxx
------------
zk = x.xxxxx
```

NOTAS:

✓ Todos los valores (excepto N) 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"
- El ***único*** código que puede ser reutilizado es el de tus programas 1 a 6
- Debe manejar apropiadamente ***todas*** las condiciones normales y anormales
- 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 = 15 wk = 185.00000 xk = 150.00000 yk = 45.00000 ------------ b0 = 0.83326 b1 = 0.04622 b2 = 0.00317 b3 = 0.18629 ------------ zk = 18.24223
Probar con una lista de datos	Teclear en pantalla: Arch2.txt	N = 8 wk = 167.00000 xk = 56.00000 yk = 155.00000 ------------ b0 = 1.81952 b1 = 0.07900 b2 = 0.03926 b3 = 0.04536 ------------ zk = 24.24150
Probar con una lista de datos	Teclear en pantalla: Arch3.txt	N = 6 wk = 650.00000 xk = 3000.00000 yk = 155.00000 ------------ b0 = 6.70134 b1 = 0.07837 b2 = 0.01504 b3 = 0.24606 ------------ zk = 140.90199

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)

Multiple regression

Overview

Suppose that you had the following data on six projects.

- development hours required
- added, reused, and modified LOC

Suppose now that you want to estimate the hours for a new project that you judged would have 650 LOC of added code, 3,000 LOC reused code, and 155 LOC of modified code.

How would you estimate the development hours and the prediction interval?

The answer is multiple regression. Multiple regression is necessary when you have combined data for several different categories of work and you cannot separate them. You have total development times for each program, but no way to determine how many hours were spent on new development, reuse, or modification. This is a common problem since it is generally impractical to separately measure the time spent in each such category of development work.

Calculating multiple regression

1. Use the following multiple regression formula to calculate the project value.
$$z_k = \beta_0 + w_k\beta_1 + x_k\beta_2 + y_k\beta_3$$

2. Find the Beta parameters by solving the following simultaneous linear equations.
$$\beta_0 n + \beta_1\sum_{i=1}^{n} w_i + \beta_2\sum_{i=1}^{n} x_i + \beta_3\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i$$
$$\beta_0\sum_{i=1}^{n} w_i + \beta_1\sum_{i=1}^{n} w_i^2 + \beta_2\sum_{i=1}^{n} w_i x_i + \beta_3\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i z_i$$
$$\beta_0\sum_{i=1}^{n} x_i + \beta_1\sum_{i=1}^{n} w_i x_i + \beta_2\sum_{i=1}^{n} x_i^2 + \beta_3\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i z_i$$
$$\beta_0\sum_{i=1}^{n} y_i + \beta_1\sum_{i=1}^{n} w_i y_i + \beta_2\sum_{i=1}^{n} x_i y_i + \beta_3\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i z_i$$

3. Solve these equations with the standard algebraic methods that are used for simultaneous linear equations. Gaussian elimination procedure provides a structured way to do this.

A multiple regression example

A multiple regression example

In this example, we'll calculate the regression parameters and predictions range using the historical data in the table below. We'll then use the regression parameters and predictions range to estimate the hours for a new project that is judged to have 650 LOC of added code, 3,000 LOC reused code, and 155 LOC of modified code.

Prog#	Added LOC	Reused LOC	Modified LOC	Development Hours
	w	**x**	**y**	**z**
1	1,142	1,060	325	201
2	863	995	98	98
3	1,065	3,205	23	162
4	554	120	0	54
5	983	2,896	120	138
6	256	485	88	61

1. Find the Beta parameters by solving the following simultaneous linear equations.

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} w_i + \beta_2 \sum_{i=1}^{n} x_i + \beta_3 \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} z_i$$

$$\beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i^2 + \beta_2 \sum_{i=1}^{n} w_i x_i + \beta_3 \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i z_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} w_i x_i + \beta_2 \sum_{i=1}^{n} x_i^2 + \beta_3 \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i z_i$$

$$\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} w_i y_i + \beta_2 \sum_{i=1}^{n} x_i y_i + \beta_3 \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i z_i$$

2. When you calculate the values of the terms, you get the following simultaneous linear equations.

$$6\beta_0 + 4,863\beta_1 + 8,761\beta_2 + 654\beta_3 = 714$$

$$4,863\beta_0 + 4,521,899\beta_1 + 8,519,938\beta_2 + 620,707\beta_3 = 667,832$$

$$8,761\beta_0 + 8,519,938\beta_1 + 21,022,091\beta_2 + 905,925\beta_3 = 1,265,493$$

$$654\beta_0 + 620,707\beta_1 + 905,925\beta_2 + 137,902\beta_3 = 100,583$$

3. Diagonalize, using Gauss' method. This successively eliminates one parameter at a time from the equations by successive multiplication and subtraction to give the following.

$$6\beta_0 + 4,863\beta_1 + 8,761\beta_2 + 654\beta_3 = 714$$

$$0\beta_0 + 580,437.5\beta_1 + 1,419,148\beta_2 + 90,640\beta_3 = 89,135$$

$$0\beta_0 + 0\beta_1 + 4,759,809\beta_2 - 270,635\beta_3 = 5,002.332$$

$$0\beta_0 + 0\beta_1 + 0\beta_2 + 37,073.93\beta_3 = 9,122.275$$

Continued on next page

A multiple regression example, continued

4. Solve for the Beta terms.

$\beta_0 = 6.7013$

$\beta_1 = 0.0784$

$\beta_2 = 0.0150$

$\beta_3 = 0.2461$

5. The final estimate is as follows.

z = 6.71+0.0784*650+0.0150*3,000+0.2461*155
= 140.902 hours

The gauss method

The gauss method

A frequently occurring problem in scientific engineering and business applications is to solve a system of N equations with N unknowns.

Gaussian elimination is a commonly used algorithm for solving these equations.

The Gauss method converts this

$$\boxed{\begin{array}{l} 2X + 3Y = 8 \\ 3X + 6Y = 15 \end{array}}$$

to this

$$\boxed{\begin{array}{l} X + 2Y = 5 \\ Y = 2 \end{array}}$$

This technique is called *triangularization.*

What is the value of X?

Gauss method example

Solve this system.

The pivot [Row 1, Col 1]

$$\boxed{\begin{array}{l} 2X + 3Y = 8 \\ 3X + 6Y = 15 \end{array}}$$

1. The equation with the largest absolute coefficient in the pivot column [any row, col 1] is exchanged with the equation in the pivot row.

$$\boxed{\begin{array}{l} 3X + 6Y = 15 \\ 2X + 3Y = 8 \end{array}}$$

2. Reduce the coefficient for the pivot term to 1 by multiplying through by the reciprocal of the coefficient.

$$\boxed{\begin{array}{l} (1/3 * 3X) + (1/3 * 6Y) = (1/3 * 15) \\ 2X + 3Y = 8 \end{array}}$$

We now have

$$\boxed{\begin{array}{l} x + 2Y = 5 \\ 2X + 3Y = 8 \end{array}}$$

Continued on next page

The gauss method, continued

3. Eliminate the pivot column terms in all rows below the pivot row
3a. Multiply the pivot row by the negated coefficient of the row below the pivot term to be eliminated, -2 in this example.

$$\boxed{\begin{array}{l} X + 2Y = 5 \\ 2X + 3Y = 8 \end{array}} \qquad \boxed{-2} * \boxed{X + 2Y = 5} = \boxed{-2X + -4Y = -10}$$

Note: don't change the pivot row; put this result in a temp store.

3b. Add the 3a result to the row being eliminated.
3. Repeat 3a and 3b for each row below the pivot term.

$$\boxed{\begin{array}{l} X + 2Y = 5 \\ 2X + 3Y = 8 \end{array}} + \boxed{-2X + -4Y = -10} = \boxed{\begin{array}{l} X + 2Y = 5 \\ -Y = -2 \end{array}} \quad \boxed{Y = 2}$$

4. Complete the triangularization by repeating steps 1 to 3 as needed, then solve for the remaining terms using back-substitution.

$$\boxed{\begin{array}{l} x + 2y = 5 \\ Y = 2 \end{array}}$$

$$\boxed{\begin{array}{l} X + 4 = 5 \\ Y = 2 \end{array}}$$

$$\boxed{\begin{array}{l} X = ? \\ Y = 2 \end{array}}$$

Continued on next page

The gauss algorithm

On a computer, the equations are represented by a matrix of the coefficients.

X	Y	
2	3	8
3	6	15

The algorithm is the same.

Given

X	Y	
2	3	8
3	6	15

1.

X	Y	
3	6	15
2	3	8

2.

X	Y	
1	2	5
2	3	8

3.

X	Y	
1	2	5
0	-1	-2

X	Y	
1	2	5
0	1	2

4.

X	Y	
1	2	5
0	1	2

X=1
Y=2

Continued on next page

A gauss algorithm

1. $I = 1$

2. LET $P = A_{K,I} = \max\{|A_{J,I}| : I <= J <= N\}$ *(Find the pivot value which is the largest number in the column below the I,I position.)*

3. IF $P = 0$ THEN EXIT (*If the remainder of this column is zero, a unique solution does not exist.)*

4. IF $K > I$ THEN FOR $J = I$ TO $N + 1$ SWAP A(I,J) AND A(K,J) *(Get the pivot value into the I,I position.)*

5. FOR $J = I + 1$ TO $N + 1$ LET $A_{I,J} = A_{I,J} / A_{I,I}$ *(Get a one in the I,I position.)* SET $A_{I,I} = 1$.

6. FOR $L = I + 1$ to N multiply row I by $-A_{L,I}$ and add to row L *(Zero out the column below the I,I position.)*

7. $I = I + 1$. IF $I < N$ THEN GO TO 2.

8. EXIT to back-substitution