

# Programación Genética: La Regresión Simbólica

## Resumen:

*Si se tiene una serie de valores históricos. ¿Se puede predecir el comportamiento futuro? La respuesta es: muy probable. ¿Y cómo? Analizando esos valores históricos por si hay algún patrón que los conecta. Una vez encontrado el patrón ya se puede hacer predicciones realistas. ¿Y cómo se halla ese patrón? Existe un procedimiento estadístico llamado correlación entre variables, pero, no funciona en todos los casos. ¿Qué otros procedimientos existen para hallar patrones cuando la correlación falla? Existe uno llamado la Regresión Simbólica que es una aplicación de la Programación Genética (una rama de la Inteligencia Artificial).*

Se tiene el siguiente problema: ¿Cómo se comportará el mercado en los próximos meses? Eso trasnocha a todo Gerente porque si supiera que va a pasar, entonces podría tomar medidas para maximizar la rentabilidad de la empresa o minimizar las pérdidas. Saber que sucederá en el futuro es sueño de todo ser humano.

Afortunadamente existe una herramienta estadística que puede colaborarnos para visualizar como se comportará el futuro si poseemos datos históricos. Se le conoce como Análisis de Regresión.

Por ejemplo, si se tiene los siguientes datos históricos

Mes	Valor
1	3
2	5
3	7
4	9
5	11

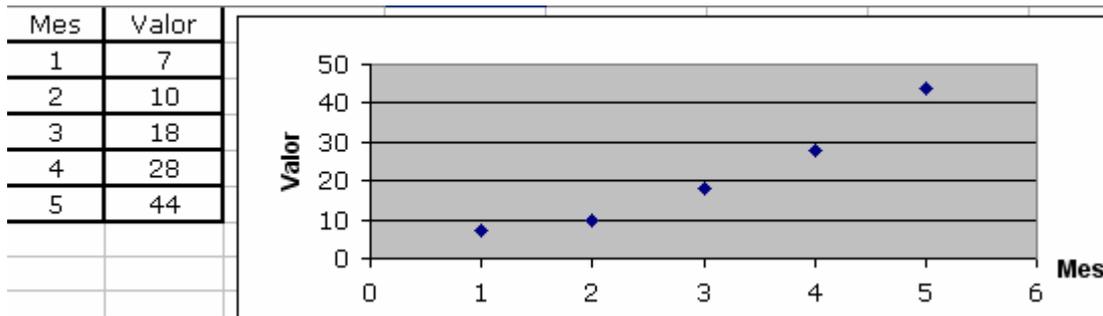
Se puede deducir que la ecuación es: **Valor = 2\*Mes+1** y con esa ecuación se puede predecir que en el mes 6 el valor será 13.

En el ejemplo anterior, es relativamente sencillo deducir la ecuación, pero si se enfrenta a datos históricos como por ejemplo:

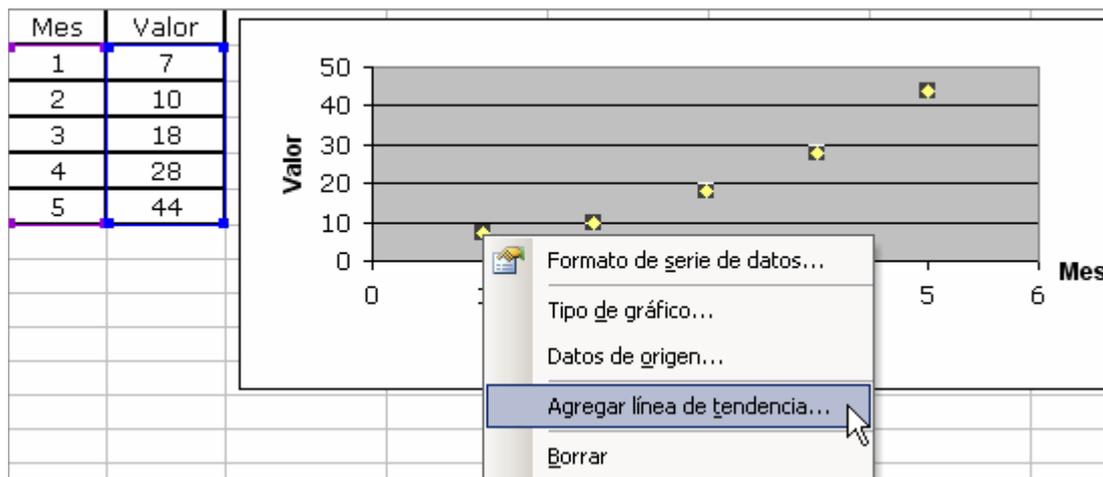
Mes	Valor
1	7
2	10
3	18
4	28
5	44

Ya no es tan sencillo dar con una ecuación. El análisis de regresión nos permite encontrar ecuaciones que se aproximen a esos valores.

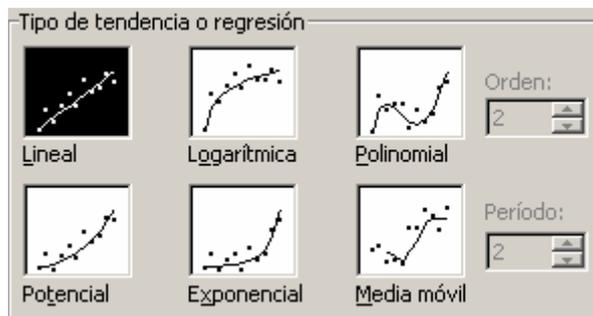
Utilizando una hoja de cálculo (como Microsoft Excel) simplemente se escriben los datos como se ve en la imagen y se genera la gráfica estadística.



Se genera la línea o curvas de tendencia para poder dar con la mejor ecuación que se acerque a los datos históricos.



Hay varias curvas para escoger: lineal, exponencial, polinomial, etc. en el ejemplo se selecciona "Lineal"



Y en las opciones se selecciona "Presentar ecuación en el gráfico" y "Presentar el valor R cuadrado en el gráfico"

**Agregar línea de tendencia**

Tipo Opciones

Nombre de la línea de tendencia

Automática: Lineal (Serie1)

Personalizada:

Extrapolar

Hacia delante: 0 Unidades

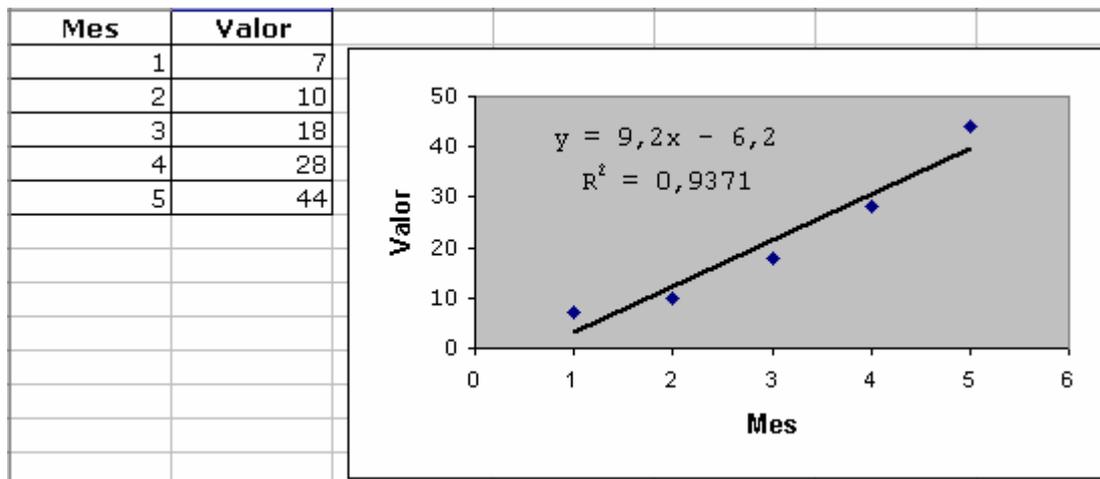
Hacia atrás: 0 Unidades

Señalar intersección = 0

Presentar ecuación en el gráfico

Presentar el valor R cuadrado en el gráfico

Este es el resultado obtenido



La hoja de cálculo dedujo que si es una línea de tendencia recta, la ecuación con mejor aproximación es  $y = 9,2x - 6,2$  con una aproximación  $R^2 = 0,9371$

El valor  $R^2$  se conoce como coeficiente de correlación lineal o correlación de Pearson, el cual muestra la exactitud de la correlación: si tiene valor 1 la correlación es perfecta, por encima de 0,65 se considera buena, entre 0,4 y 0,649 se considera regular y por debajo de este valor se considera mala. Lo ideal, por lo tanto, es que sea de valor 1. Esta correlación se calcula así:

$Y_t$  (tendencia): Es el resultado de reemplazar X en la ecuación deducida por la hoja de cálculo.

$X$	$Y_e$	$Y_t$ (tendencia)	$Y_e - \bar{Y}_e$	$Y_t - \bar{Y}_t$	$(Y_e - \bar{Y}_e) * (Y_t - \bar{Y}_t)$
1	7	3	-14,4	-18,4	264,96
2	10	12,2	-11,4	-9,2	104,88
3	18	21,4	-3,4	0	0
4	28	30,6	6,6	9,2	60,72
5	44	39,8	22,6	18,4	415,84
					<b>846,4 (sumatoria)</b>

$$\bar{Y}_e = 21,4 \text{ (promedio)}$$

$$\bar{Y}_t = 21,4 \text{ (promedio)}$$

Calculamos la covarianza:

$$Covarianza = \frac{\sum (Y_e - \bar{Y}_e) * (Y_t - \bar{Y}_t)}{n - 1} = \frac{846,4}{5 - 1} = 211,6$$

$$S_{ye} = \text{Desviación típica}$$

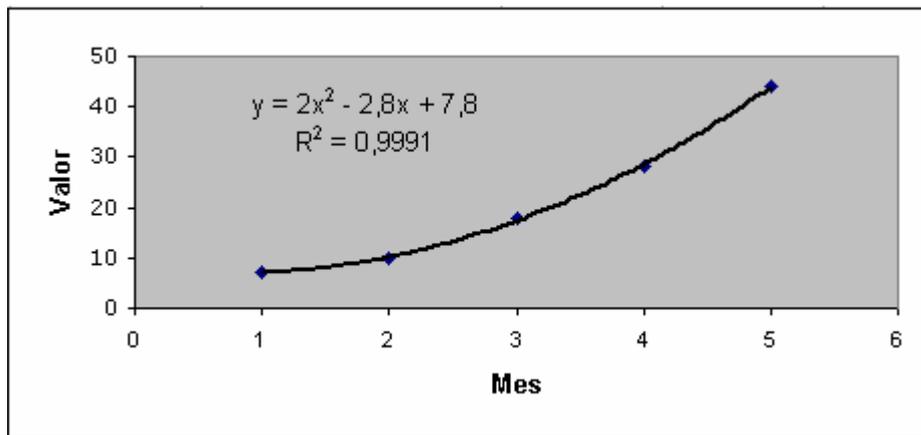
$$S_{ye} = 15,02664301$$

$$S_{yt} = \text{Desviación típica de } Y_t = 14,54647724$$

$$r = \frac{Covarianza}{S_{ye} * S_{yt}} = \frac{211,6}{15,02664301 * 14,54647724} = 0,9680457$$

Correlación de Pearson =  $r^2 = 0,9680457^2 = 0,9371125$  que es el valor que nos muestra automáticamente la hoja de cálculo.

Si seleccionamos otra ecuación, como la polinomial de grado 2 se obtiene una ecuación que es mejor como se ve en la imagen. Inclusive la correlación de Pearson está muy cercana a 1.



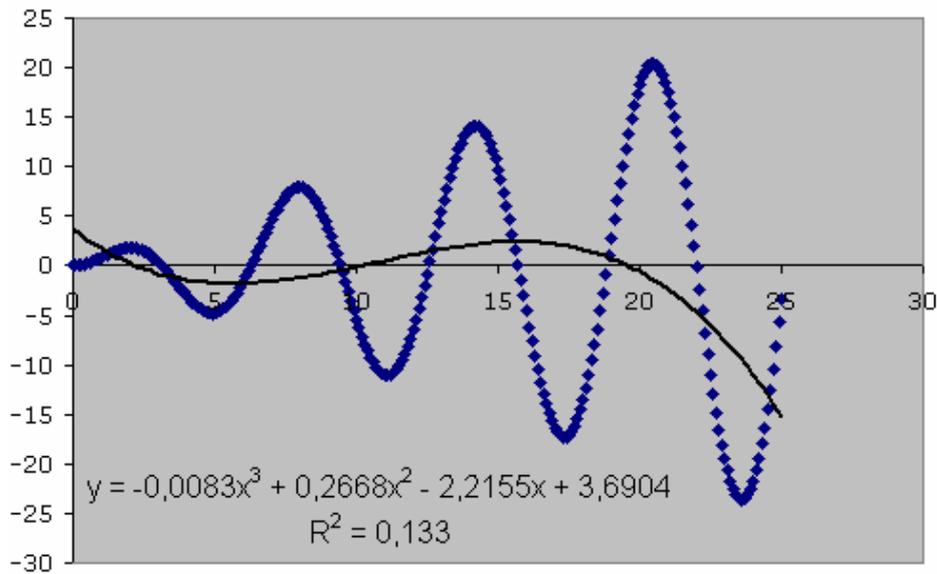
Una segunda técnica conocida como suma de diferencias absolutas puede ser usada para verificar la calidad de las ecuaciones (entre mas se acerque al valor cero o sea cero mejor).

Mes	Valor Y	Ecuación 1	Diferencia absoluta entre Ecuación 1 y Valor Y	Ecuación 2	Diferencia absoluta entre Ecuación 2 y Valor Y
1	7	3	4	7	0
2	10	12,2	2,2	10,2	0,2
3	18	21,4	3,4	17,4	0,6
4	28	30,6	2,6	28,6	0,6
5	44	39,8	4,2	43,8	0,2
Sumando las diferencias			16,4		1,6

Efectivamente, en este caso, la ecuación polinomial de grado 2 es más precisa que la ecuación lineal y por lo tanto, es mejor usar esa ecuación para predecir el futuro (por ejemplo para saber como será el sexto mes, solo ingrese el valor 6 a X).

La labor del usuario, es entonces encontrar la mejor ecuación (idealmente que la correlación de Pearson sea uno(1) o la suma de diferencias absolutas sea cero).

Sin embargo, encontrar la mejor ecuación para la siguiente serie de puntos mostrada no es sencillo.



Ecuación de Línea de Tendencia	Correlación de Pearson
$y = -0,2402x + 2,0044$	0,0291
$y = -0,0458x^2 + 0,9038x - 2,7433$	0,0734
$y = -0,0083x^3 + 0,2668x^2 - 2,2155x + 3,6904$	0,133

La hoja de cálculo nos lleva hasta un valor polinomial de grado 6, pero su correlación de Pearson es **0,195**, en otras palabras, no es posible con esta herramienta encontrar una correlación adecuada.

### ***Una aplicación de la Programación Genética para hallar las ecuaciones de correlación: La Regresión Simbólica.***

La Regresión Simbólica se define como la búsqueda de funciones que se ajusten a una serie de puntos dado. La Regresión Simbólica tiene acceso a una mayor variedad de funciones (por ejemplo, funciones sinusoidales, valor absoluto, división modular, truncamiento, redondeos, polinomios de grado N, raíces, etc.) comparada con la búsqueda determinística mostrada en la hoja de cálculo que está limitada a trabajar con los siguientes tipos de tendencia: polinomio (de grado 2 a grado 6), lineal, logarítmica, potencial y exponencial.

Estos son los pasos que deben darse para hacer Regresión Simbólica (inspirados en como trabaja la naturaleza y la evolución de las especies<sup>2</sup>):

1. Se genera una población inicial de N ecuaciones (formadas al azar).
2. A cada ecuación se le deduce su correlación de Pearson.
3. Ordene las ecuaciones en orden descendente (según el dato de la correlación de Pearson) y seleccione las M primeras.
4. Esas M ecuaciones son reproducidas y generan Q hijas, pero cada hija al azar se le hace un pequeño cambio (esto se conoce como mutación) o la hija es el producto combinado de dos ecuaciones padres (esto se conoce como cruce).

5. Reemplace las Q ecuaciones peores de la población inicial con las Q ecuaciones hijas.
6. Vuelva al punto 2 repitiéndose el ciclo constantemente hasta que se obtengan ecuaciones con un valor de correlación de Pearson aceptable.

Ejemplo: Dada la siguiente serie de puntos

X	Ye
0	0
0,1	0,00998334
0,2	0,03973387
0,3	0,08865606
0,4	0,15576734
0,5	0,23971277
0,6	0,33878548
0,7	0,45095238
0,8	0,57388487
0,9	0,70499422
1	0,84147098

Determinar la mejor ecuación  $f(x)$  que dado el valor de X se acerque a Ye

Paso 1: Se genera una serie de ecuaciones al azar (Población=N=10)

$Y = 2 * \text{seno}(x) + \text{coseno}(x)$
$Y = 5 * \text{tangente}(x/3) - \text{coseno}(x/2)$
$Y = 17 - \text{seno}(x) + \text{coseno}(x)$
$Y = 4.91 - \text{tangente}(x) - \text{coseno}(x) + \text{seno}(x)$
$Y = x * \text{coseno}(x) - \text{seno}(x) + 7.19$
$Y = x * x - \text{seno}(x/2) + 3$
$Y = 2 * x - 4 * \text{seno}(x) + 6 * \text{coseno}(x)$
$Y = 1 - \text{seno}(2 * x) + \text{tangente}(x) - \text{seno}(x) - \text{tangente}(x)$
$Y = 2 * \text{seno}(x) - \text{coseno}(x * x)$
$Y = 5 - \text{seno}(x * 2) - \text{coseno}(x - x) + \text{tangente}(4)$

Paso 2: Se deduce a cada ecuación su correlación de Pearson

Ecuación	Correlación de Pearson
$Y = 2 * \text{seno}(x) + \text{coseno}(x)$	0,945542986
$Y = 5 * \text{tangente}(x/3) - \text{coseno}(x/2)$	0,960871256
$Y = 17 - \text{seno}(x) + \text{coseno}(x)$	0,964631111
$Y = 4,91 - \text{tangente}(x) - \text{coseno}(x) + \text{seno}(x)$	0,510068073
$Y = x * \text{coseno}(x) - \text{seno}(x) + 7,19$	0,961439088
$Y = x * x - \text{seno}(x/2) + 3$	0,924849655
$Y = 2 * x - 4 * \text{seno}(x) + 6 * \text{coseno}(x)$	0,985267147
$Y = 1 - \text{seno}(2 * x) + \text{tangente}(x) - \text{seno}(x) - \text{tangente}(x)$	0,771527756
$Y = 2 * \text{seno}(x) - \text{coseno}(x * x)$	0,963507321
$Y = 5 - \text{seno}(x * 2) - \text{coseno}(x - x) + \text{tangente}(4)$	0,629947087

Paso 3: Orden en forma descendente y tome los tres primeros puestos (M=3)

<b>Ecuación</b>	<b>Correlación de Pearson</b>
<b><math>Y = 2*x - 4*\text{seno}(x) + 6*\text{coseno}(x)</math></b>	<b>0,985267147</b>
<b><math>Y = 17-\text{seno}(x)+\text{coseno}(x)</math></b>	<b>0,964631111</b>
<b><math>Y = 2*\text{seno}(x)-\text{coseno}(x*x)</math></b>	<b>0,963507321</b>
$Y = x*\text{coseno}(x)-\text{seno}(x) + 7,19$	0,961439088
$Y = 5*\text{tangente}(x/3)-\text{coseno}(x/2)$	0,960871256
$Y = 2*\text{seno}(x)+\text{coseno}(x)$	0,945542986
$Y = x*x - \text{seno}(x/2) + 3$	0,924849655
$Y = 1-\text{seno}(2*x)+\text{tangente}(x)-\text{seno}(x)-\text{tangente}(x)$	0,771527756
$Y = 5-\text{seno}(x*2)-\text{coseno}(x-x)+\text{tangente}(4)$	0,629947087
$Y = 4,91-\text{tangente}(x)-\text{coseno}(x)+\text{seno}(x)$	0,510068073

Paso 4: Tome esas M ecuaciones padres y genere Q hijas (las hijas varían por mutación). Observe el cambio de padre a hija. En este caso cada padre genera una sola hija (Q=3)

<b>Ecuación Padre</b>	<b>Ecuación Hija</b>
$Y = 2*x - 4*\text{seno}(x) + 6*\text{coseno}(x)$	$Y = x - 4*\text{seno}(x) + 6*\text{coseno}(x)$
$Y = 17-\text{seno}(x)+\text{coseno}(x)$	$Y = \text{seno}(x)+\text{coseno}(x)$
$Y = 2*\text{seno}(x)-\text{coseno}(x*x)$	$Y = 2*\text{coseno}(x)-\text{coseno}(x*x)$

Paso 5: Reemplace las peores ecuaciones de la población anterior con las hijas

<b>Ecuación</b>	<b>Correlación de Pearson</b>
$Y = 2*x - 4*\text{seno}(x) + 6*\text{coseno}(x)$	0,985267147
$Y = 17-\text{seno}(x)+\text{coseno}(x)$	0,964631111
$Y = 2*\text{seno}(x)-\text{coseno}(x*x)$	0,963507321
$Y = x*\text{coseno}(x)-\text{seno}(x) + 7,19$	0,961439088
$Y = 5*\text{tangente}(x/3)-\text{coseno}(x/2)$	0,960871256
$Y = 2*\text{seno}(x)+\text{coseno}(x)$	0,945542986
$Y = x*x - \text{seno}(x/2) + 3$	0,924849655
$Y = x - 4*\text{seno}(x) + 6*\text{coseno}(x)$	
$Y = \text{seno}(x)+\text{coseno}(x)$	
$Y = 2*\text{coseno}(x)-\text{coseno}(x*x)$	

Paso 6: Vuelva al Paso 2 hasta que se obtenga una ecuación cuya correlación de Pearson sea aceptable (por ejemplo mas de 0,9988)

No hay un absoluto que así siempre se deba proceder en hacer regresión simbólica utilizando programación genética, el concepto se mantiene pero los pasos pueden tener variaciones:

1. En el ejemplo se trabajó con una población inicial de 10, pero también puede usarse 100, 1000 o 2000 (una población grande daría mas variedad pero también requiere mas poder de computo).
2. Las ecuaciones generadas usan funciones como seno, coseno, tangente, pero podría haberse usado otras funciones como exponenciales, logaritmo natural, cosecante, etc. (da una mayor variedad).
3. La función de evaluación fue la correlación de Pearson pero esta requiere varios cálculos por ecuación para conseguirla (lo que conlleva a usar mucho poder de computo), quizás sea mejor usar otro cálculo más rápido como la suma de diferencias absolutas.
4. La selección de las mejores ecuaciones para ser reproducidas fue simplemente ordenar de mayor a menor y extraer las primeras M ecuaciones; el número a escoger puede variar y la escogencia puede ser de otra manera (por torneo, por ruleta, etc.) siempre premiando las mejores pero no cerrándole la puerta de tajo a las que no les fue tan bien.
5. Solo se generó una hija por padre. Pueden generarse dos o mas ecuaciones hijas, además solo se usó el operador de mutación para variar la hija del padre, también se puede hacer cruce.
6. Las ecuaciones hijas reemplazaron las peores ecuaciones, pero también puede usarse otra forma de reemplazo (por ejemplo, se escogen Q ecuaciones de la población en forma aleatoria y esas serán reemplazadas por las hijas).

### ***Problemas con el uso de la Programación Genética***

1. Requiere alto poder de cómputo.
2. Se debe pensar como formar las ecuaciones al azar de tal manera que haya una gran variedad pero que sean algebraicamente correctas.
3. El resultado de este procedimiento depende mucho de cómo se implementa el algoritmo y los parámetros a usar (población inicial, criterio de selección, número de hijos, reemplazo, etc.)
4. Consume gran cantidad de números aleatorios (el generador de números aleatorios debe tener un periodo bastante grande como el Mersenne Twister).
5. La simulación puede que quede paralizada en un máximo local (La población se quedó formada con ecuaciones que por mas mutaciones o cruces se hagan, no hay forma de mejorar).
6. Las ecuaciones generadas no están factorizadas e inclusive pueden tener operaciones que se anulan entre si (por ejemplo:  $x-x+x-x$ ).
7. No hay garantías que la ecuación encontrada sea la mejor.

### ***Conclusiones***

Toda teoría científica al final genera una aplicación práctica, la Programación Genética y su aplicación en Regresión Simbólica es una aplicación de la Teoría de la Evolución de las Especies de Charles Darwin. Solo es una pequeña muestra de lo que se puede hacer con lo descubierto por Darwin.

La Regresión Simbólica nos permite ir más allá de la simple relación causal entre dos variables, la realidad trabaja con muchas variables. Obsérvese la siguiente tabla:

Valor X	Valor Y	Valor Z
2	7	<b>15</b>

4	2	<b>18</b>
5	1	<b>9</b>
7	5	<b>11</b>

Valor Z es el resultado de una ecuación en que participan el Valor X y el Valor Y. La pregunta es ¿cuál es la ecuación?, ¿será  $Z=3*X-2*Y$  o  $Z=5*X*X+X*Y-3*Y$  o ....? ¿Si nos enfrentamos a problemas con tres, cuatro, cinco o más variables involucradas?

La vida es la respuesta a los problemas complejos de múltiples variables que nos plantea la realidad y la evolución es la herramienta que la vida usa para resolverlos. ¿Qué nos impide hacer uso de esta herramienta para resolver nuestros problemas?

## ***Bibliografía***

JOSÉ SANTOS, RICHARD J. DURO. Evolución Artificial y Robótica Autónoma. Alfaomega, Ra-Ma. 2005.

NILS J. NILSSON. Inteligencia Artificial. Una nueva síntesis. McGrawHill. 2001.

ANTONIO VÉLEZ. Del Big Bang al Homo sapiens. Villegas Editores. 2004

CHARLES DARWIN. El Origen de las Especies. 1859

## ***Citas***

<sup>1</sup> En Microsoft Excel, la desviación típica o desviación estándar se calcula con la función =DESVEST(rango)

<sup>2</sup> La Programación Genética tiene como base la Teoría de la Evolución de las Especies de Charles Darwin



Rafael Alberto Moreno Parra

Ingeniero de Sistemas. Universidad de San Buenaventura, Cali. Docente hora cátedra, facultad de Ingeniería de Sistemas, Universidad Libre. Docente hora cátedra, programa de Ingeniería de Sistemas e Ingeniería Industrial, Universidad de San Buenaventura.

Página Web de Investigación sobre Vida Artificial:  
<http://darwin.50webs.com>