# Programación genética: La regresión simbólica

Rafael Alberto Moreno Parra


## Resumen

El análisis de regresión es una técnica estadística que busca deducir el patrón de una serie de datos o investigar la relación estadística entre una variable dependiente (Y) y una o más variables independientes, el resultado es una expresión algebraica del tipo Y=F(X1, X2, ...Xn). En este artículo se trabaja con el tipo de análisis de regresión más usual: la regresión lineal que tiene una variable independiente Y=F(X). El usuario común tiene contacto con la regresión lineal al usar las hojas electrónicas que implementen la deducción de líneas de tendencia dada una serie de datos. Sin embargo, se percatará que existen varios límites en esta técnica, por ejemplo, los datos tienen comportamientos sinusoidales o siguen un comportamiento de alguna función algebraica o combinación de funciones algebraicas por fuera del menú ofrecido: lineal, polinomial, potencial, logarítmica o exponencial. La regresión simbólica (una aplicación de la Programación Genética) tiene el mismo objetivo de la regresión lineal pero con un espectro mucho mayor de búsqueda y menos limitaciones: Dados los datos, buscará el patrón (expresión algebraica) que identifique el comportamiento de estos accediendo a todo tipo de funciones y combinaciones algebraicas.



## Abstract

***Regression analysis is a statistical analysis that aims to deduct the pattern in a series of data or research the statistical relation between a dependent variable (Y) and one or more dependent variables, the result is an algebraic expression type Y=F (X1, X2, …Xn). This article has the most common regression analysis: lineal regression which has one independent variable Y=F(X). A common user comes into contact with lineal regression when using electronic sheets that implement tendency line deduction given a series of data. However, he/she will notice there are certain limits to this technique for example, the data has sinusoidal behavior or follows some algebraic function behavior or a combination of algebraic functions beyond the offered menu: lineal, polynomial, potential, logarithmic or exponential. Symbolic regression (a genetic programming application) has the same objective as lineal regression but with a much greater search spectrum and much less limitations: Given the data, it will search for the pattern (algebraic expression) that identifies their behavior ascending to all types of functions and algebraic combinations.***



## Palabras clave

***Programación genética, regresión simbólica, análisis de regresión, inteligencia artificial, evolución artificial, computación evolutiva.***

## Keywords

***Genetic programming, symbolic regression, regression analysis, artificial intelligence, artificial evolution, evolutionary computation***


SISTEMAS



Fecha de recepción: 01 - 05- 2007 Fecha de aceptación: 13 - 06 - 2007

Entramado Vol.3 No. 1, 2007 (Enero - Junio)

La regresión simbólica (una aplicación de la Programación Genética) tiene el mismo objetivo de la regresión lineal pero con un espectro mucho mayor de búsqueda y muchas menos limitaciones. Dados los datos, buscará la expresión algebraica que identifique el comportamiento de estos accediendo a todo tipo de funciones y combinaciones algebraicas.

## Introducción

**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 pueden 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).**

## Planteamiento del problema

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

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

**Por ejemplo, si se tienen 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.**

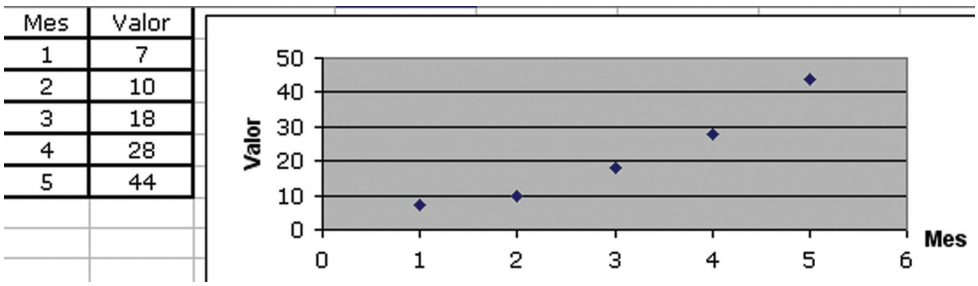
**En el ejemplo anterior, es relativamente sencillo deducir la ecuación, pero si se enfrenta a datos históricos como por ejemplo:**

| 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.**

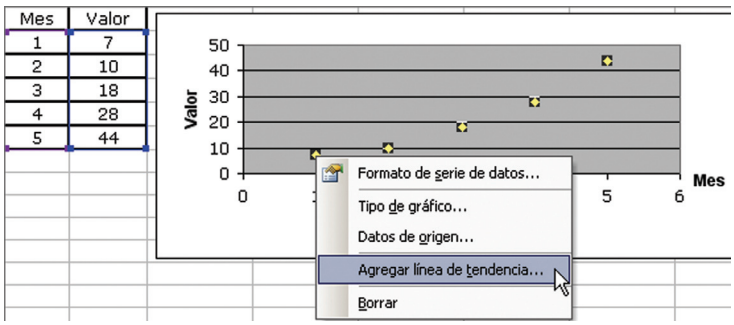
## Análisis de Regresión

**Utilizando una hoja de cálculo (como Microsoft Excel[1]) simplemente se escriben los datos como se ve en la imagen y se genera una gráfica estadística (Ver Gráfica 1).**



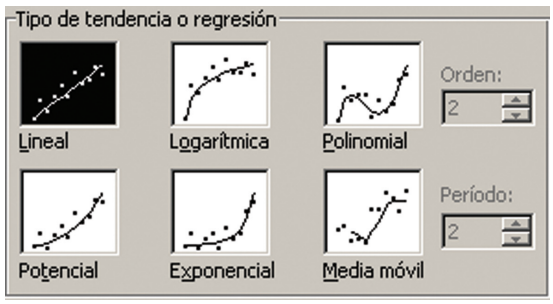

▲ Gráfica 1. Datos dados para generar una gráfica estadística con una hoja de cálculo (como Microsoft Excel)

**Se genera la línea o curvas de tendencia para poder dar con la mejor ecuación que se acerque a los datos históricos (Ver Gráfica estadística 2).**



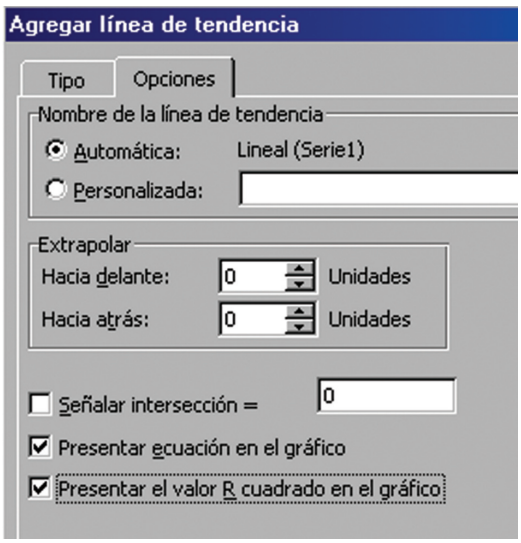

▲ Gráfica 2. Línea o curva de tendencia generada

**Hay varias curvas para escoger: lineal, exponencial, polinomial, etc. en el ejemplo se selecciona "Lineal" (Ver Gráfica estadística 3).**




▲ Gráfica 3. Tipos de curvas para escoger

**Por último en las opciones se selecciona** "Presentar ecuación en el gráfico" y "Presentar el valor R cuadrado en el gráfico" **(Ver Gráfica 4).**




▲ Gráfica 4. Indicaciones para agregar línea de tendencia

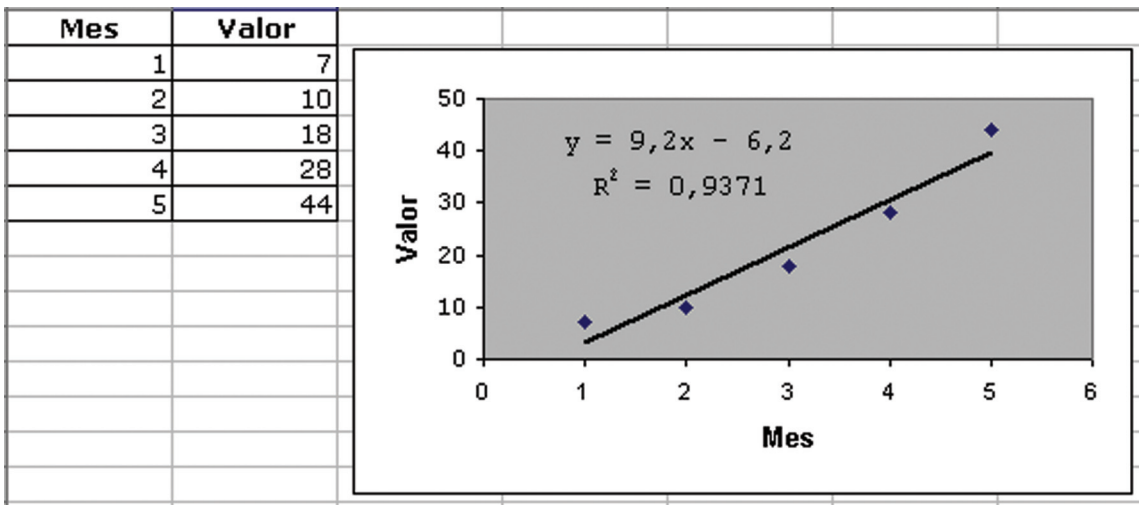
**El resultado obtenido se puede oservar en la Gráfica 5 (Ver gráfica 5).**

**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.




▲ Gráfica 5. Resultado final: Gráfica generada o curva de tendencia generada

**La 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) | $Y_e - \overline{Y}_e$ | $Y_t - \overline{Y}_t$ | $(Y_e - \overline{Y}_e * Y_t - \overline{Y}_t)$ |
|---|---|---|---|---|---|
| 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 |
| | | | | | **846,40** |

$\overline{Y}_e$ = **21,4 (promedio)**

$\overline{Y}_t$ = **21,4 (promedio)**

**Calculamos la covarianza:**

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

$S_{ye}$ = **Desviación típica**

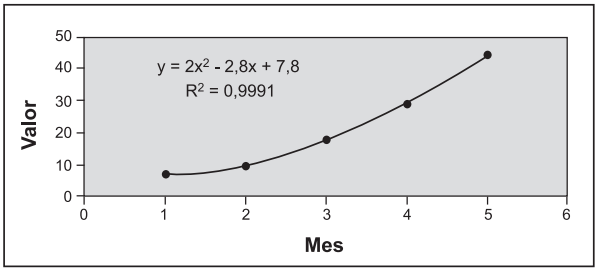
$_{1\ de}\ Y_e$ = **15,02664301**

$S_{ye}$ = **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 como la que se observa en la Gráfica 6, que es mejor. Inclusive la correlación de Pearson está muy cercana a 1 (Ver Gráfica 6).**




▲ Gráfica 6: Ecuación polinomial de grado 2

**Una segunda técnica conocida como suma de diferencias absolutas puede ser usada para verificar la calidad de las ecuaciones (entre más se acerque al valor cero o sea cero mejor) (Ver Tabla 1).**

| 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** |

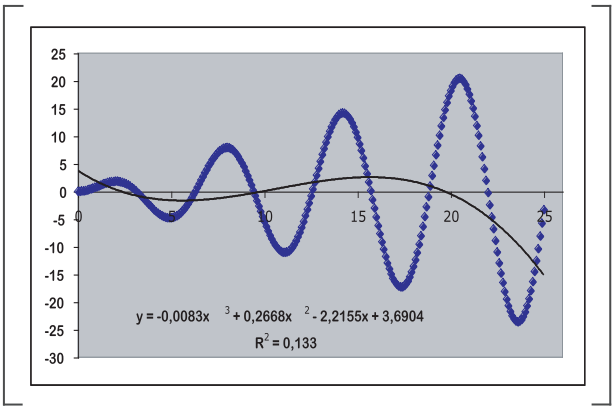
▲ Tabla 1: Suma de diferencias

**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 cómo 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 serie de puntos que se observa en la Gráfica 7 no es sencillo.**

**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. (Ver Tabla 2)**




▲ Gráfica 7: Ecuación polinomial de grado 2

| Ecuación de línea de tendencia | Correlación de Pearson |
|---|---|
| y = -0,2402x + 2,0044 | 0,0291 |
| y = -0,0458x*x + 0,9038x - 2,7433 | 0,0734 |
| y = -0,0083x*x*x + 0,2668x*x - 2,2155x + 3,6904 | 0,133 |

▲ Tabla 2: Ecuación polinomial de grado

## 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 cómo trabaja la naturaleza y la evolución de las especies[2]):**

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 | $Y_e$ |
|---|---|
| 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)**

| Ecuaciónes generadas |
|---|
| Y = 2*seno(x)+coseno(x) |
| Y = 5*tangente(x/3)-coseno(x/2) |
| Y = 17-seno(x)+coseno(x) |
| Y = 4.91-tangente(x)-coseno(x)+seno(x) |
| Y = x*coseno(x)-seno(x) + 7.19 |
| Y = x*x – seno(x/2) + 3 |
| Y = 2*x – 4*seno(x) + 6*coseno(x) |
| Y = 1-seno(2*x)+tangente(x)-seno(x)-tangente(x) |
| Y = 2*seno(x)-coseno(x*x) |
| Y = 5-seno(x*2)-coseno(x-x)+tangente(4) |

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

| Ecuación | Correlación de Pearson |
|---|---|
| Y = 2*seno(x)+coseno(x) | 0,945542986 |
| Y = 5*tangente(x/3)-coseno(x/2) | 0,960871256 |
| Y = 17-seno(x)+coseno(x) | 0,964631111 |
| Y = 4.91-tangente(x)-coseno(x)+seno(x) | 0,510068073 |
| Y = x*coseno(x)-seno(x) + 7.19 | 0,961439088 |
| Y = x*x – seno(x/2) + 3 | 0,924849655 |
| Y = 2*x – 4*seno(x) + 6*coseno(x) | 0,985267147 |
| Y = 1-seno(2*x)+tangente(x)-seno(x)-tangente(x) | 0,771527756 |
| Y = 2*seno(x)-coseno(x*x) | 0,963507321 |
| Y = 5-seno(x*2)-coseno(x-x)+tangente(4) | 0,629947087 |

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

| Ecuación | Correlación de Pearson |
|---|---|
| **Y = 2*x – 4*seno(x) + 6*coseno(x)** | **0,985267147** |
| **Y = 17-seno(x)+coseno(x)** | **0,964631111** |
| **Y = 2*seno(x)-coseno(x*x)** | **0,963507321** |
| Y = x*coseno(x)-seno(x) + 7.19 | 0,961439088 |
| Y = 5*tangente(x/3)-coseno(x/2) | 0,960871256 |
| Y = 2*seno(x)+coseno(x) | 0,945542986 |
| Y = x*x – seno(x/2) + 3 | 0,924849655 |
| Y = 1-seno(2*x)+tangente(x)-seno(x)-tangente(x) | 0,771527756 |
| Y = 5-seno(x*2)-coseno(x-x)+tangente(4) | 0,629947087 |
| Y = 4.91-tangente(x)-coseno(x)+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)**

| Ecuación padre | Ecuación hija |
|---|---|
| Y = 2*x – 4*seno(x) + 6*coseno(x) | Y = x – 4*seno(x) + 6*coseno(x) |
| Y = 17-seno(x)+coseno(x) | Y = seno(x)+coseno(x) |
| Y = 2*seno(x)-coseno(x*x) | Y = 2*coseno(x)-coseno(x*x) |

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

| Ecuación | Correlación de Pearson |
|---|---|
| **Y = 2*x – 4*seno(x) + 6*coseno(x)** | **0,985267147** |
| **Y = 17-seno(x)+coseno(x)** | **0,964631111** |
| **Y = 2*seno(x)-coseno(x*x)** | **0,963507321** |
| Y = x*coseno(x)-seno(x) + 7.19 | 0,961439088 |
| Y = 5*tangente(x/3)-coseno(x/2) | 0,960871256 |
| Y = 2*seno(x)+coseno(x) | 0,945542986 |
| Y = x*x – seno(x/2) + 3 | 0,924849655 |
| Y = x – 4*seno(x) + 6*coseno(x) | |
| Y = seno(x)+coseno(x) | |
| Y = 2*coseno(x)-coseno(x*x) | |

**Paso 6: Vuelva al Paso 2 hasta que se obtenga una ecuación cuya correlación de Pearson sea aceptable (por ejemplo, más 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 más variedad pero también requiere más poder de cómputo).
2. Las ecuaciones generadas usan funciones como seno, coseno, tangente, pero podrían 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 usar mucho poder de cómputo), 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 cerrándoles la puerta de tajo a las que no les fue tan bien.
5. Solo se generó una hija por padre. Pueden generarse dos o más 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 cómo 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 más 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 sí (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 | 15 |
| 4 | 2 | 18 |
| 5 | 1 | 9 |
| 7 | 5 | 11 |

**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?**

# CITAS

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

2 La programación genética tiene como base la Teoría de la Evolución de las Especies de Charles Darwin

# BIBLIOGRAFÍA

**SANTOS J., Richard J. Duro. Evolución Artificial y Robótica Autónoma. Alfaomega, Ra-Ma. 2005.**

**NILSSON Nils, J. . Inteligencia Artificial. Una nueva síntesis. McGrawHill. 2001.**

**VÉLEZ, Antonio . Del Big Bang al Homo sapiens. Villegas Editores. 2004**

**DARWIN, Charles . El Origen de las Especies. 1859**

**Rafael Alberto Moreno Parra**

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