The VirtualDynamics Intelligencer
Journal of Research and Development

ISSN: 2955-8387 (on line)
Vol 1(1) July 2022

Hiperajuste (HP) o Ajuste de Regresión Lineal Ideal

Galo Emilio Sisniegas Charcape
gsisniegasch@yahoo.com

Resumen.

El método de regresión lineal es sumamente útil para correlacionar dos magnitudes que se supone podrían tener una funcionalidad lineal. Ahora, dadas dos magnitudes, digamos X y Y. Sí a X se designa como la variable independiente, Y será la dependiente, entonces, se tratará como $Y\ vs. X$, pero si a X se toma como dependiente la Y es la independiente, entonces el tratamiento será $X\ vs. Y$. Ahora, al hacer el ajuste de rectas para el caso $Y\ vs. X$ se encuentra que al expresar la recta solución en su forma general se expresaría como $A + BX + CY = 0$, pero cuando tratamos el caso $X\ vs. Y$ resultará en $D + EX + FY = 0$, una ecuación general completamente distinta, entonces, aunque podría ser poca la diferencia de todos modos resulta en soluciones ambiguas. Así, en este artículo se plantea y desarrolla un método de regresión lineal que, aunque resulta, también, en dos soluciones ninguna resulta ambigua. Una de ellas se interpretaría como el ajuste lineal deseado, mientras que la otra solución podría interpretarse como la ecuación de ajuste ortogonal correspondiente. Para el desarrollo del método nos fundamentamos en el concepto físico de *equilibrio de torques* o *momentos de fuerza* tomando a un conjunto de puntos como un sistema que no rota, por tanto, la suma de torques se asume como cero e identificando el punto de giro definimos los dos elementos fundamentales para la aplicación de la idea de torque, el análogo al brazo de palanca y el análogo de la fuerza aplicada.

Palabras Clave: Regresión lineal. Ajuste lineal. Ajuste de recta. Recta ajustada. Dispersión. Centro de masa. Gravicentro. Torque. Simetría.

Introducción.

El autor desarrolló el método aquí expuesto en 1997 [1], mientras era profesor en la Facultad de Ciencias Físicas de la Universidad Nacional Mayor de San Marcos (Lima – Peru), aunque desarrollado originalmente de modo escalar y no vectorial como el presentado en el presente artículo.

Consideremos un conjunto de puntos cuyas coordenadas muestran una aparente relación lineal tal como presentamos en la figura 1, se ha elegido puntos que muestran un cierto grado de dispersión. Por cuestiones de visualización, los ejemplos de las figuras muestran un conjunto con un número reducido de puntos.

Planteamiento

Consideremos $D_i = (X_i, Y_i)$, donde $i = 1,2, \dots, n$, un punto determinado de un conjunto de datos, constituido de **n** puntos cuyos pares coordenados son

$$(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots, (X_n, Y_n) \quad (1)$$

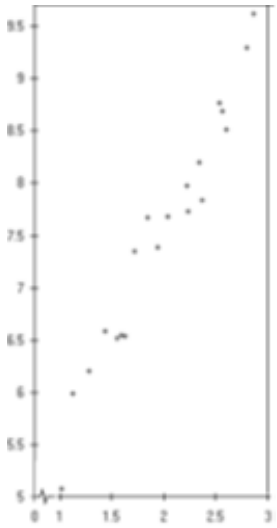

Figura 1. Se muestra un diagrama de dispersión de un conjunto de datos (puntos).

Se busca la línea recta ajustada a los datos (ver la figura.2) que representaremos por

$$y = a + bx \quad (2)$$

donde se desea encontrar la pendiente b y el intercepto a. También, definamos una línea recta perpendicular a la línea recta buscada, ecuación (2), que cruza al *punto dato* D_i dado y cuya ecuación es la siguiente

$$y = a^i - \frac{1}{b}x \quad (3)$$

donde $-1/b$ y a^i son la pendiente y el intercepto respectivamente, el valor del intercepto depende a su vez del *punto dato* dado D_i del conjunto de puntos (1).

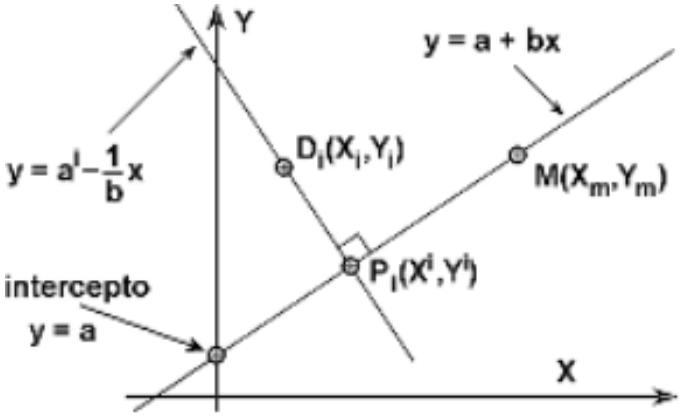

Figura 2. Se muestra el "centro de masa" M ("gravicentro") del conjunto, un punto dato dado D_i, la línea recta ajustada buscada $y = a + bx$, y el punto P_i correspondiente al punto de intersección de la línea recta ajustada y la línea recta perpendicular que cruza al punto dato D_i del conjunto.

Se define ahora el punto $P_i = (X^i, Y^i)$ como el punto de intersección de las líneas rectas definidas por la ecuación (2) y la ecuación (3). Éste es el punto más cercano a la recta ajustada desde el punto dato dado D_i, también, se debe considerar a $M = (X_m, Y_m)$ como el punto que

determina el "centro de masa" (“gravicentro”) de todo el conjunto de datos punto definido en la (1), ver figura 2.

Se puede expresar los interceptos a y a^i como funciones de la pendiente b con ayuda de las coordenadas de los puntos M y D_i utilizando respectivamente la ecuación (2) y la ecuación (3), así

$$Y_m = a + bX_m \quad \Rightarrow \quad a = Y_m - bX_m \tag{4}$$

y a partir de ecuación (3), se tiene

$$Y_i = a^i - \frac{1}{b}X_i \quad \Rightarrow \quad a^i = \frac{bY_i + X_i}{b} \tag{5}$$

Ya que la ordenada del punto intersección P_i es la misma para la ecuación (2) y la ecuación (3), entonces

$$bX^i + a = -\frac{1}{b}X^i + a^i$$

y despejando la abscisa X^i y reemplazando los interceptos con ayuda de la ecuación (4) y la ecuación (5) se obtiene

$$X^i = \frac{b(Y_i - Y_m) + X_i + b^2X_m}{b^2 + 1} \tag{6}$$

y para la ordenada correspondiente

$$Y^i = \frac{b^2Y_i + Y_m + b(X_i - X_m)}{b^2 + 1} \tag{7}$$

entonces las diferencias entre las coordenadas del punto D_i y del punto P_i, dependerán sólo de b, son las siguientes

$$\begin{aligned} X_i - X^i &= \frac{b^2(X_i - X_m) - b(Y_i - Y_m)}{b^2 + 1} \\ Y_i - Y^i &= \frac{-b(X_i - X_m) + (Y_i - Y_m)}{b^2 + 1} \end{aligned} \tag{8}$$

también las ecuaciones para las diferencias entre las coordenadas del punto P_i y del punto M dependerán sólo de b, así,

$$\Rightarrow \quad \begin{aligned} X^i - X_m &= \frac{b(Y_i - Y_m) + (X_i - X_m)}{b^2 + 1} \\ Y^i - Y_m &= \frac{b^2(Y_i - Y_m) + b(X_i - X_m)}{b^2 + 1} \end{aligned} \tag{9}$$

De esta forma es posible expresar los vectores correspondientes $\overrightarrow{MP_i}$ y $\overrightarrow{P_iD_i}$ como a continuación:

$$\begin{aligned} \overrightarrow{MP_i} &= \langle X^i - X_m, Y^i - Y_m, 0 \rangle \\ \overrightarrow{P_iD_i} &= \langle X_i - X^i, Y_i - Y^i, 0 \rangle \end{aligned} \tag{10}$$

y detallando las coordenadas de las ecuaciones (10) con las ecuaciones (8) y las ecuaciones (9) se tiene

$$\overrightarrow{MP_i} = \left\langle \frac{b(Y_i - Y_m) + (X_i - X_m)}{b^2 + 1}, \frac{b^2(Y_i - Y_m) + b(X_i - X_m)}{b^2 + 1}, 0 \right\rangle \tag{11}$$

$$\overrightarrow{P_iD_i} = \left\langle \frac{b^2(X_i - X_m) - b(Y_i - Y_m)}{b^2 + 1}, \frac{-b(X_i - X_m) + (Y_i - Y_m)}{b^2 + 1}, 0 \right\rangle \tag{12}$$

y ahora estamos preparados para la aplicación del concepto de torque.

Aplicación del concepto de torque

Con la ecuación (11) y la ecuación (12) es posible definir el "*torque"* [4], $\vec{T_i}$, del punto D_i con respecto al "punto eje de giro" M (figura 2). Entonces

$$\vec{T_i} = \overrightarrow{P_iD_i} \times \overrightarrow{MP_i} \tag{13}$$

donde los vectores $\overrightarrow{MP_i}$ y $\overrightarrow{P_iD_i}$ corresponderían respectivamente con el brazo del torque y la fuerza del torque. Entonces efectuando la operación

$$\vec{T_i} = \begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ \dfrac{b^2(X_i - X_m) - b(Y_i - Y_m)}{b^2+1} & \dfrac{-b(X_i - X_m) + (Y_i - Y_m)}{b^2+1} & 0 \\ \dfrac{b(Y_i - Y_m) + (X_i - X_m)}{b^2+1} & \dfrac{b^2(Y_i - Y_m) + b(X_i - X_m)}{b^2+1} & 0 \end{vmatrix}$$

ó

$$\vec{T_i} = \frac{1}{b^2+1}\begin{vmatrix} \hat{\imath} & \hat{\jmath} & \hat{k} \\ b^2(X_i - X_m) - b(Y_i - Y_m) & -b(X_i - X_m) + (Y_i - Y_m) & 0 \\ b(Y_i - Y_m) + (X_i - X_m) & b^2(Y_i - Y_m) + b(X_i - X_m) & 0 \end{vmatrix}$$

$$\vec{T_i} = \frac{\langle 0,0,(b^4-1)(X_i - X_m)(Y_i - Y_m) + b(b^2+1)[(X_i - X_m)^2 - (Y_i - Y_m)^2]\rangle}{b^2+1}$$

Es decir,

$$\vec{T_i} = \langle 0,0,(b^2-1)(X_i - X_m)(Y_i - Y_m) + b[(X_i - X_m)^2 - (Y_i - Y_m)^2]\rangle$$

Que pasamos a representar como

$$\vec{T_i} = \langle 0,0,(b^2-1)\kappa_i\gamma_i + b({\kappa_i}^2 - {\gamma_i}^2)\rangle \tag{14}$$

donde

$$\kappa_{Xi} = X_i - X_m \quad y \quad \kappa_{Yi} = Y_i - Y_m \tag{15}$$

cuyo módulo sería

$$T_i = (b^2-1)\kappa_{Xi}\kappa_{Yi} + b({\kappa_{Xi}}^2 - {\kappa_{Yi}}^2) \tag{16}$$

Equilibrando los "torques"

De esta manera es posible hacer la sumatoria de los "torques" correspondientes a cada punto del conjunto de los datos punto. Si consideramos que los puntos del conjunto de datos, (1), se distribuirían alrededor de una hipotética línea recta ajustada de modo que habría un equilibrio de torques, entonces de acuerdo con el principio de equilibrio de torques la sumatoria será igual a cero. Así

$$\sum_{i=1}^{n} T_i = 0 \tag{17}$$

Así,

$$\sum_{i=1}^{n}[(b^2-1)\kappa_{Xi}\kappa_{Yi} + b({\kappa_{Xi}}^2 - {\kappa_{Yi}}^2)] = 0 \tag{18}$$

y la ecuación se reduce a

$$(b^2-1)\sum_{i=1}^{n}\kappa_{Xi}\kappa_{Yi}+b\sum_{i=1}^{n}({\kappa_{Xi}}^2-{\kappa_{Yi}}^2)=0 \tag{19}$$

Luego asignando un símbolo a cada sumatoria

$$S_a=\sum_{i=1}^{n}\kappa_{Xi}\kappa_{Yi} \quad y \quad S_b=\sum_{i=1}^{n}({\kappa_{Xi}}^2-{\kappa_{Yi}}^2) \tag{20}$$

es posible reescribir la ecuación (19) como una cuadrática cuyos coeficientes son las sumatorias:

$$S_a b^2+S_b b-S_a=0 \tag{21}$$

Recuérdese que las variables κ_{Xi} y κ_{Yi} que aparecen en cada elemento de las sumatorias (20) se definen por medio de las ecuaciones (15).

Resolviendo la ecuación cuadrática

A partir de la ecuación (21) se obtiene

$$b=\frac{-S_b\pm\sqrt{(S_b)^2+4(S_a)^2}}{2S_a}$$

la cual es posible reescribir como

$$b=-S\pm\sqrt{S^2+1} \tag{22}$$

donde

$$S=\frac{S_b}{2S_a} \tag{23}$$

De este modo se obtienen dos soluciones a partir de la ecuación (22): Una es la pendiente de la línea recta ajustada al eje longitudinal y la otra al eje transversal, véase la figura 3.

Relación entre las soluciones de la cuadrática

Las dos soluciones de la ecuación (22) son:

$$b=-S+\sqrt{S^2+1} \tag{24}$$

y

$$-\frac{1}{b}=-S-\sqrt{S^2+1} \tag{25}$$

Entonces la sumatoria de miembro a miembro de estas dos ecuaciones resulta en

$$b-\frac{1}{b}=-2S \tag{26}$$

Arreglando la ecuación (26) y considerando la ecuación (23) se tiene

$$\frac{b^2-1}{b}=-\frac{S_b}{S_a} \quad \Longrightarrow \quad \frac{b^2-1}{b}=-\frac{\sum_{i=1}^{n}({\kappa_{Xi}}^2-{\kappa_{Yi}}^2)}{\sum_{i=1}^{n}\kappa_{Xi}\kappa_{Yi}} \tag{27}$$

que denominaremos la *ecuación sumatoria de soluciones*.

Ahora considerando las definiciones dadas por las ecuaciones (15) y las ecuaciones (20) obtengamos las expresiones para los promedios de las sumatorias dadas por las ecuaciones (20). Así desarrollando la sumatoria S_a:

$$S_a = \sum_{i=1}^{n} \kappa_{Xi}\kappa_{Yi} = \sum_{i=1}^{n} (X_i - X_m)(Y_i - Y_m)$$

ó

$$S_a = \sum_{i=1}^{n} (X_iY_i - X_iY_m - X_mY_i + X_mY_m) \tag{28}$$

Que promediándola resulta la siguiente expresión

$$\frac{S_a}{n} \equiv \Pi_a = (XY)_m - 2X_mY_m + X_mY_m$$

Es decir,

$$\Pi_a = (XY)_m - X_mY_m \tag{29}$$

Efectuando análogamente con la sumatoria S_b:

$$S_b = \sum_{i=1}^{n} ({\kappa_{Xi}}^2 - {\kappa_{Yi}}^2) = \sum_{i=1}^{n} [(X_i - X_m)^2 - (Y_i - Y_m)^2]$$

$$= \sum_{i=1}^{n} \left[\left({X_i}^2 - 2X_iX_m + {X_m}^2\right) - \left({Y_i}^2 - 2Y_iY_m + {Y_m}^2\right)\right]$$

Así,

$$S_b = \sum_{i=1}^{n} \left[{X_i}^2 - {Y_i}^2 + 2(Y_iY_m - X_iX_m) + {X_m}^2 - {Y_m}^2\right] \tag{30}$$

Que, promediándola, $\frac{S_b}{n}$, resulta la siguiente expresión

$$\Pi_b = (X^2)_m - (Y^2)_m + {Y_m}^2 - {X_m}^2 \tag{31}$$

Colocando la ecuación (29) y la ecuación (31) en reemplazo de las sumatorias que aparecen en la ecuación (27), y dado que

$$\frac{S_b}{S_a} = \frac{\frac{S_b}{n}}{\frac{S_a}{n}}$$

que equivale a

$$\frac{S_b}{S_a} = \frac{\Pi_b}{\Pi_a} \tag{32}$$

se obtiene la *ecuación sumatoria de soluciones* (27) expresada en función de los promedios dados en la ecuación (29) y la ecuación (31)

$$\frac{b^2 - 1}{b} = -\frac{(X^2)_m - {X_m}^2 + {Y_m}^2 - (Y^2)_m}{(XY)_m - X_mY_m}$$

Es decir,

$$\frac{b^2 - 1}{b} = \frac{(Y^2)_m - {Y_m}^2 + {X_m}^2 - (X^2)_m}{(XY)_m - X_mY_m} \tag{33}$$

Efecto de la traslación de ejes en la ecuación sumatoria de las soluciones.

Si se efectúa una traslación de ejes del sistema XY al sistema X'Y' cuyo origen está situado en el punto (x_0, y_0) según el sistema XY, las variables de la ecuación original variarán del modo siguiente:

$$X \to X' \equiv X + x_0 \quad y \quad Y \to Y' \equiv Y + y_0 \tag{34}$$

Entonces los valores promedios de las coordenadas de los datos tendrán el siguiente efecto:

$$\begin{gathered} X'_m = (X + x_0)_m = X_m + x_0 \\ y \\ Y'_m = (Y + y_0)_m = Y_m + y_0 \end{gathered} \tag{35}$$

Y en cuanto al producto de sus coordenadas se tiene que

$$\begin{gathered} (X'Y')_m = [(X + x_0)(Y + y_0)]_m = [XY + Xy_0 + x_0Y + x_0y_0]_m \\ \equiv \quad (X'Y')_m = (XY)_m + X_m y_0 + x_0 Y_m + x_0 y_0 \end{gathered} \tag{36}$$

Y a los cuadrados de sus coordenadas resultan en lo siguiente

$$\begin{gathered} (X'^2)_m = (X^2 + 2x_0X + {x_0}^2)_m = (X^2)_m + 2x_0X_m + {x_0}^2 \\ y \\ (Y'^2)_m = (Y^2 + 2y_0Y + {y_0}^2)_m = (Y^2)_m + 2y_0Y_m + {y_0}^2 \end{gathered} \tag{37}$$

Además, definamos el producto de los promedios de sus coordenadas:

$$X'_m Y'_m = (X + x_0)_m (Y + y_0)_m = X_m Y_m + X_m y_0 + x_0 Y_m + x_0 y_0 \tag{38}$$

También definamos los cuadrados de los promedios de sus coordenadas como:

$$\begin{gathered} {X'_m}^2 = (X_m + x_0)^2 = {X_m}^2 + 2x_0X_m + {x_0}^2 \\ y \\ {Y'_m}^2 = (Y_m + y_0)^2 = {Y_m}^2 + 2y_0Y_m + {y_0}^2 \end{gathered} \tag{39}$$

Ahora definamos las diferencias de miembro a miembro de los valores dados por la ecuación (36) y las ecuaciones (37), con las dadas por la ecuación (38) y las ecuaciones (39), respectivamente:

$$(X'Y')_m - X'_m Y'_m = (XY)_m - X_m Y_m \tag{40}$$

$$(X'^2)_m - {X'_m}^2 = (X^2)_m - {X_m}^2 \tag{41}$$

$$(Y'^2)_m - {Y'_m}^2 = (Y^2)_m - {Y_m}^2 \tag{42}$$

Entonces dada la siguiente razón

$$\frac{(Y'^2)_m - {Y'_m}^2 + {X'_m}^2 - (X'^2)_m}{(X'Y')_m - X'_m Y'_m} \tag{43}$$

Y considerando las identidades dadas en la ecuación (40), ecuación (41) y ecuación (42) e introduciéndolas en la ecuación (43) se obtiene que

$$\frac{(Y'^2)_m - {Y'_m}^2 + {X'_m}^2 - (X'^2)_m}{(X'Y')_m - X'_m Y'_m} = \frac{(Y^2)_m - {Y_m}^2 + {X_m}^2 - (X^2)_m}{(XY)_m - X_m Y_m} \tag{44}$$

Y comparando la ecuación (44) con la (33) se concluye que el valor dado por la *ecuación sumatoria de soluciones* **no se afecta** por una traslación.

Relación entre los interceptos de las soluciones del HP.

El valor del correspondiente intercepto estará determinado por la ecuación (4) y por la ecuación (5). Así mientras la ecuación (4) da el intercepto de una de las soluciones por

$$a = Y_m - bX_m$$

donde X_m y Y_m son los promedios para cada componente de los puntos del conjunto de datos. El intercepto de la otra solución ortogonal a la primera, considerando la ecuación (5), sería determinada por

$$\sum_{i=1}^{N} Y_i = a^i - \frac{1}{b}\sum_{i=1}^{N} X_i \tag{45}$$

y promediando se obtiene

$$Y_m = a^i - \frac{1}{b}X_m \tag{46}$$

Entonces despejando

$$a^i = Y_m + \frac{1}{b}X_m \tag{47}$$

Es el intercepto de la recta ajustada perpendicular, ecuación (3), que cruza en el punto M a la recta ajustada buscada, ecuación (2).

Sumatoria y diferencia entre los interceptos.

La suma y diferencia entre los interceptos dados por la ecuación (47) y la ecuación (4) resultan en lo siguiente:

$$a^i + a = 2Y_m + \left(\frac{1}{b} - b\right)X_m \tag{48}$$

y

$$a^i - a = \left(\frac{1}{b} + b\right)X_m \tag{49}$$

Producto y razón entre los interceptos. El producto y el cociente entre los interceptos dados por ecuación (47) y ecuación (4) resultan en lo siguiente:

$$a^i a = \left(Y_m + \frac{1}{b}X_m\right)(Y_m - bX_m)$$

$$\equiv \quad a^i a = \left({Y_m}^2 - {X_m}^2\right) + \left(\frac{1}{b} - b\right)X_m Y_m \tag{50}$$

y

$$\frac{a^i}{a} = \frac{Y_m + \frac{1}{b}X_m}{Y_m - bX_m} \equiv \quad \frac{a^i}{a} = \frac{bY_m + X_m}{b(Y_m - bX_m)} \tag{51}$$

Comparación con el método de regresión lineal.

A fin de efectuar comparaciones se consideró un conjunto de datos punto disperso y aleatoriamente distribuido, figura 1. En la figura 3(a) se puede ver el resultado obtenido por el método clásico de regresión lineal **[3,4]**.

Recordemos que cuando nos refiramos a pendiente será en relación con el ángulo que hace la línea con el eje x o de las abscisas. Así al analizar la figura 3(a), se puede apreciar que la línea de mayor pendiente se genera cuando se elige a la ordenada como variable independiente. De otro modo, cuando se elige la abscisa como variable independiente el resultado es la línea recta con menor pendiente que se muestra en la misma figura 3(a).

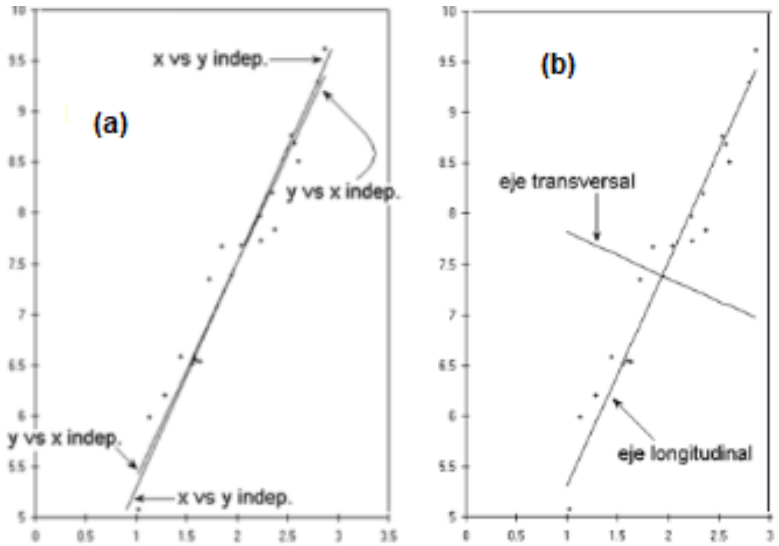

Figura 3. **(a)** Diagrama de dispersión y las soluciones, abscisa independiente y ordenada independiente, dadas por el método de regresión lineal clásico. **(b)** Se muestra las soluciones mutuamente ortogonales, la longitudinal y la transversal, obtenidas por el nuevo método, identificada por HP.

En la figura 4 se muestran los resultados de ambos métodos. La regresión clásica genera soluciones ambiguas, en cambio con el método alternativo, aunque proporciona, también, dos resultados, son de naturaleza complementaria y sin ambigüedad.

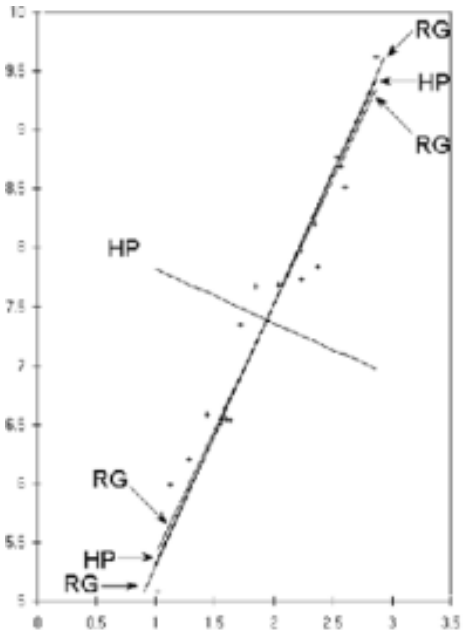

Figura 4. Se confrontan los resultados del método de regresión lineal (RG) y el del hiperajuste (HP). En este caso la solución longitudinal del método HP está próxima con una pendiente ligeramente menor a la solución por el método de RG clásico cuando se considera la ordenada como variable independiente.

Así en el método de hiperajuste, una de las soluciones tiende a ajustarse a lo largo de la distribución de puntos, a modo de un *eje longitudinal*, mientras que la otra se ajustaría a lo ancho de la misma, a modo de un *eje transversal*. Estas rectas, ya que por naturaleza están relacionadas con el concepto de torque, podrían ser denominadas *ejes de equilibrio*.

Observación: *El método que hemos denominado de Hiperajuste o HP no sería un método de mínimos cuadrados* **[3,4]**. *Por tanto, no se le puede evaluar bajo ese criterio, el cual se suele utilizar para juzgar la benignidad de cualquier método de regresión.*

Conclusiones.

El algoritmo presentado en este artículo puede utilizarse para encontrar los *ejes de equilibrio* de cualquier sistema de puntos distribuidos en un plano, a dichos ejes hemos propuesto denominarlos el *eje longitudinal* y el *eje transversal* determinables sin ambigüedad para cualquiera distribución de datos punto en un plano común.

Incluso, por ejemplo, es aplicable al conjunto de los puntos de una imagen o fotografía digitalizada, sugiriéndonos que podría usarse para analizar formas de modo que podamos evaluar su grado de simetría. En consecución, también se podría investigar la utilidad del algoritmo HP para analizar distribuciones con otras tendencias no lineales y determinar su grado de asimetricidad. En un futuro comunicado podría explicar su aplicabilidad a puntos distribuidos volumétricamente.

Referencias

[1] G. Sisniegas, Rev.Inv.Fis.9(1)78-81(2006)
[2] R. Resnick & D. Halliday, Física (parte I), Cap.13 y 14.
[3] G. Canavos (1984) Probabilidad y estadística. Cap.13.
[4] Jay Devore (2008) Probabilidad y Estadística para ingeniería y Ciencias. Cap.12.

El Autor

Galo Emilio Sisniegas Charcape ha sido profesor en varias universidades particulares en Lima, Perú, en la actualidad es profesor en la UPCH. Obtuvo su bachillerato y licenciatura en física en la UNMSM de Lima - Perú y una maestría en física aplicada en la USP en Sao Paulo – Brasil.