

Unidad 5 (Extensión)

Métodos cuantitativos de análisis gráfico

Método de cuadrados mínimos – Regresión lineal

Consideremos el caso de un conjunto de mediciones (X_i, Y_i) , con error en el valor de Y_i dado por σ_i , cuyas representaciones gráficas se muestran en la Figura 1. El objetivo de esta sección es describir el procedimiento estadístico que permite obtener la línea que mejor ajusta los datos experimentales, línea de regresión, y las incertezas asociadas en su determinación. Asimismo se desea tener un modo de estimar las incertezas asociadas a la estimación de un dado valor Y_0 , a partir de un valor X_0 .

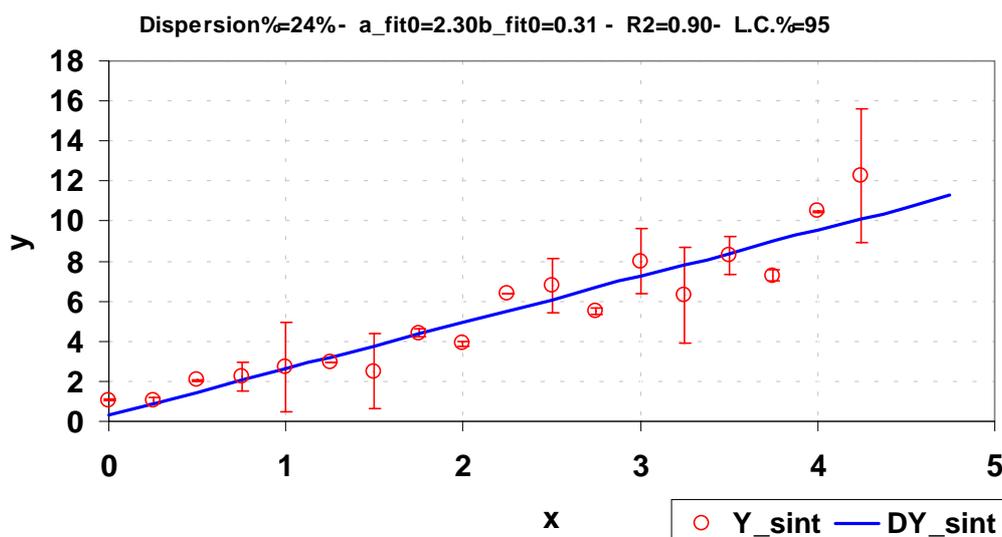


Figura 1.- Representación gráfica de un conjunto de datos experimentales (X_i, Y_i) con errores en el eje Y dado por los valores σ_i . La línea continua azul es la recta obtenida por cuadrados mínimos.

La recta que mejor ajusta los datos viene dada por la ecuación:

$$Y(x) = a \cdot x + b \quad (1)$$

Definimos el valor de Chi-Cuadrado, χ^2 , como:

$$\chi^2 = \sum_{i=1}^N w_i \cdot (Y_i - a \cdot x_i - b)^2 \quad (2)$$

Aquí W_i es un factor de peso o ponderación que se puede definir de distintos modos según el problema en estudio. Un modo usual de pesar los datos es hacerlo usando sus respectivos errores σ_j del siguiente modo:

$$w_i = \frac{1}{\sigma_i^2} \quad (3)$$

con

$$W = \sum_i w_i \quad (4)$$

Si todos los datos tienen igual ponderación, es decir si, $w_i=1$, entonces $W=N =$ número total de datos. Desde luego, la expresión (3) representa solo una de las tantas formas en que pueden ponderarse los datos. La elección más adecuada de los factores de ponderación depende del problema específico en consideración.

El método de cuadrados mínimos consiste en elegir como los mejores valores de a y b aquellos valores que minimicen el valor de χ^2 ec.(2). El resultado de este procedimiento resulta en¹⁻⁴:

$$a = \frac{W \cdot (\sum YX) - (\sum X) \cdot (\sum Y)}{W \cdot (\sum X^2) - (\sum X)^2} = \frac{\langle YX \rangle - \langle X \rangle \cdot \langle Y \rangle}{(\langle X^2 \rangle - \langle X \rangle^2)} = \frac{Cov(YX)}{S_X^2} \quad (5)$$

$$b = \frac{(\sum Y) \cdot (\sum X^2) - (\sum X) \cdot (\sum Y \cdot X)}{W \cdot (\sum X^2) - (\sum X)^2} = \frac{\langle X^2 \rangle \cdot \langle Y \rangle - \langle X \cdot Y \rangle \cdot \langle X \rangle}{(\langle X^2 \rangle - \langle X \rangle^2)} \quad (6)$$

o bien

$$b = \langle Y \rangle - a \cdot \langle X \rangle \quad (6')$$

Donde usamos la notación:

$$\sum Y \equiv \sum_{i=1}^N w_i \cdot Y_i, \quad (7)$$

$$\sum X^n \equiv \sum_{i=1}^N w_i \cdot X_i^n, \quad (8)$$

$$\sum Y \cdot X \equiv \sum_{i=1}^N w_i \cdot Y_i \cdot X_i \quad (9)$$

y así sucesivamente.

También definimos los valores medios de y y x como:

$$\bar{Y} \equiv \langle Y \rangle \equiv \sum_{i=1}^N w_i \cdot Y_i / W, \quad (10)$$

$$\bar{X} \equiv \langle X \rangle \equiv \sum_{i=1}^N w_i \cdot X_i / W, \quad (11)$$

$$\langle X^n \rangle \equiv \sum_{i=1}^N w_i \cdot X_i^n / W, \quad (12)$$

Las desviaciones estándar vienen dadas por:

$$S_x^2 \equiv \sum_{i=1}^N w_i \cdot (X_i - \bar{X})^2 / W = \left(\frac{N-1}{N} \right) \cdot \text{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2, \quad (13)$$

y

$$S_y^2 \equiv \sum_{i=1}^N w_i \cdot (Y_i - \bar{Y})^2 / W = \left(\frac{N-1}{N} \right) \cdot \text{Var}(Y) = \langle Y^2 \rangle - \langle Y \rangle^2, \quad (14)$$

Los coeficientes de correlación se definen en forma similar:

$$S_{YX} \equiv \text{Cov}(Y, X) = \sum_{i=1}^N w_i \cdot (X_i - \bar{X}) \cdot (Y_i - \bar{Y}) / W = \langle YX \rangle - \langle Y \rangle \langle X \rangle, \quad (15)$$

y

$$R \equiv \frac{\langle YX \rangle - \langle Y \rangle \langle X \rangle}{S_x \cdot S_y} = \frac{\text{Cov}(YX)}{S_x \cdot S_y}. \quad (16)$$

El error típico de estimación de Y sobre X , está relacionado con el valor de Chi-cuadrado, χ_N^2 , por:

$$\chi_N^2 = \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} = \text{Error.típico}(YX)^2 \quad (17)$$

También se define el valor de Chi-cuadrado por grados de libertad: χ_v^2 :

$$\chi_v^2 = sd^2 = \frac{N}{(N-2)} \cdot \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} = \frac{N}{N-2} \cdot \chi_N^2 \quad (18)$$

El parámetro, χ_N^2 se relaciona con la covarianza de XY , para el caso de una ajuste lineal (5) y (6), por la relación:

$$\chi_N^2 = S_y^2 - \frac{[\langle (X - \bar{X}) \cdot (Y - \bar{Y}) \rangle]^2}{S_x^2} = S_y^2 - \left[\frac{\text{Cov}(Y, X)}{S_x} \right]^2 \quad (19)$$

La variación total, $S_t = S_y$, da una medida de cómo los puntos Y_i se distribuyen alrededor del valor medio de Y . S_t se define como:

$$N \cdot S_t^2 = \sum_{i=1}^N w_i \cdot (Y_i - \bar{Y})^2 = N \cdot S_y^2 \quad (20)$$

La variación explicada, S_{inex} , mide la calidad del modelo, $Y(X_i)$, para explicar los datos observados, Y_i . Este nombre surge del hecho que $\varepsilon_i = (Y_i - Y(X_i))$ tienen una distribución estadística al azar. S_{inex} se define como

$$\sum_{i=1}^N w_i \cdot (Y_i - Y(X_i))^2 = W \cdot S_{inex}^2 = W \cdot \chi_N^2 = \frac{N-2}{N} \cdot W \cdot sd^2 \quad (21)$$

La variación explicada S_{exp} , se define por:

$$\sum_{i=1}^N w_i \cdot (Y(X_i) - \bar{Y})^2 = N \cdot S_{exp}^2 \quad (22)$$

A partir de (20), (21) y (22) se demuestra que:

$$S_t^2 = S_{exp}^2 + S_{inex}^2 \quad (23)$$

de donde tenemos la propiedad:

$$R^2 = \frac{S_{expl}^2}{S_t^2} = 1 - \frac{S_{inex}^2}{S_t^2} = 1 - \frac{\chi_N^2}{S_Y^2} \leq 1 \quad (24)$$

También tenemos que:

$$Cov(X, Y) = \left(\sum W_i \cdot (X_i - \bar{X}) \cdot (Y_i - \bar{Y}) \right) / W = R \cdot S_X \cdot S_Y \quad (25)$$

Una propiedad importante de los estimadores a y b , es que si los errores de las estimaciones, ε_i :

$$\varepsilon_i^2 = (Y_i - Y(X_i))^2, \quad (26)$$

es la siguiente, si tomamos muestras sucesivas de la población, o realizamos conjuntos independientes de mediciones, cada una tendrá un valor de a y b distintos en general. Los valores a y b tendrán una distribución estadística y sus valores medios vendrán dados por $\langle a \rangle$ y $\langle b \rangle$ y sus desviaciones estándar dadas por Δa y Δb respectivamente.

$$\Delta a_0 = \frac{\sqrt{\chi_N^2}}{\sqrt{N-2} \cdot S_X} = \frac{sd}{\sqrt{N-2} \cdot S_X} \quad (27)$$

y

$$\Delta b_0 = \frac{\sqrt{\chi_N^2}}{\sqrt{N-2} \cdot S_X} \cdot \langle X^2 \rangle = \frac{sd}{\sqrt{N-2} \cdot S_X} \cdot \langle X^2 \rangle = \Delta a_0 \cdot \langle X^2 \rangle \quad (28)$$

Si los errores ε_i tienen una distribución normal, la variable aleatoria t , definida por:

$$t = \frac{(a - \langle a \rangle)}{\Delta a_0} \cdot \sqrt{N-2} \quad (29)$$

presenta una distribución t -Student, con $N-2$ grados de libertad .

Para calcular la incerteza en la estimación de a (Δa) a partir de una muestra de tamaño N , con un límite de confianza de $P\%$, se calcula a partir de del valor t_p , que se obtiene de la distribución t-Student con $(N-2)$ grados de libertad

$$\text{Probabilidad}_t\text{-Student } (t < t_p) = P\% \quad (30)$$

Si se usa Excel® Microsoft este valor de t_p se calcula usando la función DISTR.T.INV((1-P); N-2). La incerteza Δa se calcula como:

$$\Delta a(P\%) \equiv \sigma_a = t_p \cdot \frac{sd}{\sqrt{N-2} \cdot S_x} = t_p \cdot \sqrt{\frac{\chi_N^2}{(N-2) \cdot S_x^2}} = t_p \cdot a \cdot \sqrt{\frac{(1/R^2 - 1)}{(N-2)}} \quad (31)$$

El error en b viene dado por:

$$\Delta b(P\%) = \Delta a(P\%) \cdot \sqrt{\langle X^2 \rangle} \quad (32)$$

Para estimar la incerteza asociada e una proyección de un nuevo valor, calculado para un valor no medido X_0 , obtenido usando la recta de regresión $Y_0 = a \cdot X_0 + b$, con un límite de confianza de $P\%$ tenemos dos casos distintos: a) estimación de la probabilidad que un valor individual de una muestra, asociada al valor de $X = X_0$ caiga con probabilidad $P\%$ entre $Y(X_0) - \Delta Y_{estim}$ y $Y(X_0) + \Delta Y_{estim}$.

$$\Delta Y_{estim}(X_0) = t_p \cdot sd \cdot \sqrt{\left(1 + N + (X_0 - \bar{X})^2 / (S_x^2)\right)} \quad (33)$$

b) estimación de la probabilidad que un valor medio de los valores una muestra, asociada al valor de $X = X_0$ caiga con probabilidad $P\%$ entre $Y(X_0) - \Delta Y_{estim_media}$ y $Y(X_0) + \Delta Y_{estim_media}$.

$$\Delta Y_{estim_media}(X_0) = t_p \cdot \frac{sd}{\sqrt{N-2}} \cdot \sqrt{\left(1 + (X_0 - \bar{X})^2 / (S_x^2)\right)} \quad (34)$$

Es usual indicar las *bandas de los intervalos de confianza* dadas por (34) como las bandas de confianza. También se utilizan las bandas determinadas por (33) y usualmente se las designa como *bandas de predicción*.

Para cada ordenada X_0 , los valores de $\Delta Y(X_0)$ definen dos curvas: $y(x) + \Delta Y(X_0)$ y $y(x) - \Delta Y(X_0)$ entre las cuales encontraremos el $P\%$ de los datos observados. Estas bandas definen los límites de confianza de $P\%$ para las predicciones de Y . Ver figura 2

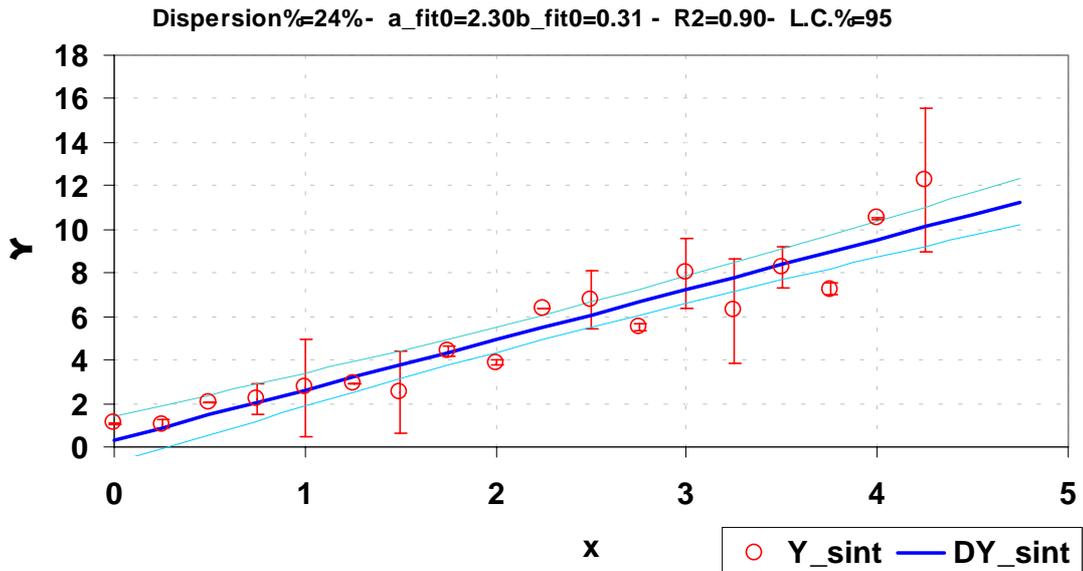


Figura 2.- Representación gráfica de un conjunto de datos experimentales (X_i, Y_i) . Las bandas laterales representan los límites de confianza de 95% (*bandas de confianza*).

Caso de Error en ambas variables: En general las técnicas estadísticas para considerar estos casos es motivo discusión entre los distintos autores y expertos en este tema. Aquí proponemos un esquema aproximado, basado fundamentalmente en las ref.4-6.

Si las mediciones (x_i, y_i) tiene errores: Δx_i y Δy_i receptivamente. Definimos los factores de peso para cada punto como:

$$W_i = 1 / \sigma_i^2 \tag{35}$$

donde:

$$\sigma_i^2 = a^2 \cdot \Delta x_i^2 + \Delta y_i^2. \tag{36}$$

En general si los factores de ponderación de la variable x e y son $w_{x,i}$ y $w_{y,i}$ respectivamente, entonces:

$$w_i = \frac{w_{x,i} \cdot w_{y,i}}{a^2 \cdot w_{y,i} + w_{x,i}} \tag{37}$$

donde a es la pendiente de la recta de regresión, ec.(5). El problema es que para determinar a debemos de resolver el problema de regresión. Para ello necesitamos los factores de peso W_i , que a su vez dependen de a . Para resolver este problema podemos proceder de modo iterativo. Usamos como ponderación inicial solo los valores de $w_{y,i}(=1/\Delta y_i^2)$, Con estos coeficientes, usando (5) obtenemos el valor de a , con este valor calculamos los pesos w_i usando (36) determinamos de nuevo los coeficientes w_i y a partir de la ec.(5) los nuevos coeficientes a . Iterando hasta que los sucesivos valores de a no cambien, se

obtienen los parámetros de la regresión lineal buscada, o sea la regresión lineal para el caso de datos con errores en las dos variables.

Estas ideas pueden extenderse al caso no lineal, en que la función $f(x;a,b,c,\dots)$ cuyos parámetros, a, b, c, \dots se buscan determinar depende de un modo no lineal de x . En este caso la generalización de (34) conduce al concepto de error efectivo⁶:

$$\sigma_i^2 = \left(\frac{df}{dx} \right)^2 \cdot \Delta x_i^2 + \Delta y_i^2. \quad (38)$$

Análogamente, la expresión (36) para los pesos o ponderación se puede generalizar en:

$$w_i = \frac{w_{x,i} \cdot w_{y,i}}{(df/dx)^2 \cdot w_{y,i} + w_{x,i}}. \quad (39)$$

Bibliografía

1. P. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, 2nd ed. (McGraw Hill, New York, 1993).
2. Stuart L. Meyer, *Data analysis for scientists and engineers* (John Willey & Sons, Inc., New York, 1975).
3. D. C. Baird, *Experimentación*, 2^a ed. (Prentice-Hall Hispanoamericana S.A., México, 1991).
4. J. Higbie, "Uncertainty in the linear regression slope" *Am. J. Phys.* **59**, 184 (1991)
5. J. Orear, "Least squares when both variables have uncertainties", *Am. J. Phys. ibid.*, **50**, 912 (1982).
6. "Simple method for fitting data when both variables have uncertainties" D. Barker and L.M. Diana *Am. J. Phys.* **42**, 224 (1974).
7. "Linear least-squares fits with errors in both coordinates." II: Comments on parameter variances - B. Cameron Reed - *Am. J. Phys.*, Vol. 60, No. 1, 1992
8. *Estadística* – M. Spiegel – McGraw Hill 2da. Ed. Bogotá 1997

Apéndice

Funciones de Excel – Incluidas en la planilla de cálculo de Excel [error_sg2k1.xls](#), en el módulo Regression_sg, escrito en Visual Basic for Application se definieron las funciones que se describen a continuación y que fueron definidas en el texto de esta introducción. En dichos programas las dimensiones de los vectores son de 300, si se trabaja con vectores de dimensiones mayores, se deben modificar dichos programas. A estas funciones *Física re-Creativa* -S. Gil y E. Rodríguez - Prentice Hall - Buenos Aires 2001

se accede a través del botón pegar funciones, sección: definidas por el usuario. Si se desea ver el código en Visual Basic, el menú de herramientas de Excel, elegir Macro: editor de Visual Basic. La mayoría de las funciones tiene incluida la opción *Mode* que se define a continuación:

<i>mode=1</i>	<i>No Weight</i>	$w_i=1$
<i>mode=2</i>	<i>Weight</i>	$w_i=1/\Delta Y_i^2$
<i>mode=3</i>	<i>Weight</i>	$w_i=1/Abs(Y_i)$
<i>mode=4</i>	<i>Weight</i>	$w_i=1/Y_i^2$
<i>mode=5</i>	<i>Weight</i>	$w_i=abs(\Delta Y_i)$

$$W = \sum_i w_i \quad (A1)$$

En general suponemos que los datos son resultados de N mediciones representados por las ternas $(X_i, Y_i, \Delta y_i)$ las cuaternas $(X_i, Y_i, \Delta y_i, \Delta x_i)$, donde los N datos X_i representan la variable independiente, representados genéricamente por el vector $Xdat$. Los errores asociados a la variable independiente, viene dados por los valores Δx_i , genéricamente representado por el vector Dx_dat . De modo análogo se definen N datos Y_i que representan la variable dependiente, representados genéricamente por el vector $Ydat$. Los errores asociados a esta variable dependiente, los designamos Δy_i , genéricamente los representamos por el vector DY_dat . Usando estas convenciones definimos las siguientes funciones:

$a_Lfit(Xdat, Ydat, DYdat, mode)$ = Pendiente de la recta de regresión ec.(5):

$$a_Lfit = a = \frac{W \cdot (\sum YX) - (\sum X) \cdot (\sum Y)}{W \cdot (\sum X^2) - (\sum X)^2} = \frac{Cov(YX)}{S_x^2} \quad (A2)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *pendiente* de Excel.

$b_Lfit(Xdat, Ydat, DYdat, mode)$ = Ordenada en el origen de la recta de regresión ec.(6):

$$b_Lfit = b = \langle Y \rangle - a \cdot \langle X \rangle \quad (A3)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *Intesección.eje* de Excel.

$Da_Lfit(Xdat, Ydat, DYdat, mode)$ = Error en la pendiente de la recta de regresión ec.(5) y (29):

$$\Delta a = Da_Lfit = \frac{sd}{\sqrt{N-2} \cdot S_x} \quad (A5)$$

Db_Lfit(Xdat, Ydat, DYdat, mode) Error en la pendiente de la recta de regresión ec.(6) y (30):

$$\Delta b = Db_Lfit = b = \Delta a \cdot \langle X^2 \rangle \quad (A6)$$

Prom_pesado_n(Ydat, DYdat, nn, mode)

$$Prom_pesado_n(X) \equiv \langle X^n \rangle = \frac{\sum_i w_i \cdot X_i^n}{W} \quad (A7)$$

Varianza(Ydat, DYdat, mode)

$$Varianza(Y) = \frac{N}{(N-1)} \cdot \frac{\sum_i w_i \cdot (Y_i - \bar{Y})^2}{W} \quad (A8)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *var* de Excel.

SYX_corr (Ydat, Yfit, DYdat, mode)

$$SYX_corr = \sqrt{\left(\frac{n}{n-2}\right) \cdot \frac{\sum_i w_i \cdot (y_i - y(x_i))^2}{W}} = \sqrt{\left(\frac{N}{N-2}\right) \cdot \left(S_Y^2 - \frac{S_{YX}^2}{S_X^2}\right)} \quad (A9)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *error.tipico. YX* de Excel.

SYnXm (Ydat, Xdat, DYdat, nn, mm, mode)

$$SYnXm = \frac{\sum_i w_i \cdot (y_i^n \cdot x_i^m)}{W} \quad (A10)$$

Prom_pesado_n(Ydat, DYdat, nn, mode)

$$SYnXm = \frac{\sum_i w_i \cdot (y_i^n \cdot x_i^m)}{W} \quad (A11)$$

R_lin (Xdat, Ydat, DYdat, mode)

$$R_lin \equiv \frac{\langle YX \rangle - \langle Y \rangle \langle X \rangle}{S_X \cdot S_Y} = \frac{S_{YX}}{S_X \cdot S_Y} \quad (A12)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *Coficiente.de.correl* de Excel.

regress_2(Ydat, Yfit, DYdat, mode)

$$regress_2 = R^2 = 1 - \left(\frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} \right) / (\langle Y^2 \rangle - \langle Y \rangle^2) \quad (A13)$$

Cuando todos los datos tienen igual peso, esta función coincide con la función *Coeficiente.R2* de Excel.

Function Chi_un (Ydat, Yfit, DYdat, mode, Nparam)

$$Chi_Nu = \left(\frac{N}{N - N_param} \right) \cdot \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} \quad (A14)$$

Chi2_tot (Ydat, Yfit, DYdat, mode)

$$Chi2_tot = N \cdot \frac{\sum w_i \cdot (Y_i - Y(X_i))^2}{W} \quad (A15)$$

Covar_YX (X_datos, Y_Datos, DY_dat, mode)

$$Cor_YX \equiv Cov(Y, X) = \sum_{i=1}^N w_i \cdot (X_i - \bar{X}) \cdot (Y_i - \bar{Y}) / W = \langle YX \rangle - \langle Y \rangle \langle X \rangle, \quad (A16)$$

SYnXm (Ydat, Xdat, DYdat, nn, mm, mode)

$$SYnXm = \sum_{i=1}^N w_i \cdot X_i^n \cdot Y_i^m / W = \langle Y^n X^m \rangle, \quad (A17)$$

Dy_lim_confianza(X_new, tp, Xdat, Ydat, DYdat, mode)

$$Dy_lim_confianza(X, tp, X_i, Y_i, \Delta Y_i, Mode) = tp \cdot \frac{sd}{\sqrt{N-2}} \cdot \sqrt{1 + \frac{(X - \bar{X})^2}{N \cdot S_X^2}}, \quad (A18)$$

dy_estima_media(X_new, tp, Chi_n, N, X_bar, Var(X))

$$dy_estima_media(X, tp, \chi_{Ni}^2, N, \bar{X}, Var(X)) = tp \cdot \frac{\sqrt{\chi_N^2}}{\sqrt{N-2}} \cdot \sqrt{1 + \frac{(X - \bar{X})^2}{S_X^2}}, \quad (A19)$$

dy_estimacion(X_new, tp, Chi_n, N, X_bar, Var(X))

$$dy_estimacion(X, tp, \chi_{Ni}^2, N, \bar{X}, Var(X)) = tp \cdot \sqrt{\chi_N^2} \cdot \sqrt{1 + N + \frac{(X - \bar{X})^2}{S_X^2}}, \quad (A20)$$

Slop_1d(Xdat, Ydat, W_dat)

$$Slop_1d = a = \frac{W \cdot (\sum YX) - (\sum X) \cdot (\sum Y)}{W \cdot (\sum X^2) - (\sum X)^2} = \frac{Cov(YX)}{S_x^2} \quad (A21)$$

Slop_2D(Xdat, Ydat, Wy_dat, Wx_dat)

$$Slop_2d = a = \frac{W \cdot (\sum YX) - (\sum X) \cdot (\sum Y)}{W \cdot (\sum X^2) - (\sum X)^2} \quad (A22)$$

La rutina que calcula la pendiente en este caso, realiza 10 iteraciones como las descritas en el texto.