Resumen de comandos sobre REGRESIÓN LINEAL con wxMaxima

Introducción de datos: Para hacer Estadística en 2D los datos deben expresarse como una *matriz de 2 columnas*, de la forma

$$\mathtt{z} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

esto puede hacerse de varios modos

Modo 1: introduciendo los datos por parejas

$$\mathtt{z} : \mathtt{matrix}([\mathtt{x_1}, \mathtt{y_1}], [\mathtt{x_2}, \mathtt{y_2}], \ldots, [\mathtt{x_n}, \mathtt{y_n}])$$

Modo 2 (recomendado): introduciendo separadamente `x` e `y`, y creando después la matriz `z`

$$\mathtt{x} : [\mathtt{x_1}, \mathtt{x_2}, \ldots, \mathtt{x_n}]; \quad \mathtt{y} : [\mathtt{y_1}, \mathtt{y_2}, \ldots, \mathtt{y_n}]$$
$$\mathtt{z} : \mathtt{transpose}(\mathtt{matrix}(\mathtt{x}, \mathtt{y}))$$

Representación gráfica: para el diagrama de puntos

`draw2d(points(z))` o bien `scatterplot(z)`

Recta de regresión: se calcula directamente con

`simple_linear_regression(z)`

devuelve la ecuación de la recta $Y = a+bX$, el coeficiente de correlación r, y alguna información adicional (ver el ejemplo).

Gráfica conjunta de diagrama de puntos y recta de regresión:

$$\mathtt{wxdraw2d}(\,\mathtt{points}(\mathtt{z}),\ \mathtt{explicit}(\mathtt{a} + \mathtt{bX}, \mathtt{X}, \mathtt{p}, \mathtt{q})\,)$$

donde '$a + bX$' es la ecuación de la recta, y $[p, q]$ un rango **adecuado** en el eje x.

Media y desviación típica: se pueden aplicar a la matriz `z`

`mean(z)`	devuelve	$[\bar{\mathtt{x}},\ \bar{\mathtt{y}}]$
`std(z)`	devuelve	$[\sigma_x,\ \sigma_y]$

Covarianza y correlación

$$\mathtt{cov}(\mathtt{z}) \quad \text{devuelve} \quad \begin{bmatrix} \sigma_x^2 & \mathrm{cov}_{x,y} \\ \mathrm{cov}_{x,y} & \sigma_y^2 \end{bmatrix}; \qquad \mathtt{cor}(\mathtt{z}) \quad \text{devuelve} \quad \begin{bmatrix} 1 & \mathrm{r}_{x,y} \\ \mathrm{r}_{x,y} & 1 \end{bmatrix}$$

NOTAS:

1.- `wxdraw2d` permite añadir algunos adornos, por ejemplo

```
wxdraw2d(explicit(3+2*X,X,-2,2),title="ejerc 2",xlabel="peso",ylabel="altura")
```

Ver más sobre decoración en *http://riotorto.users.sourceforge.net/gnuplot/func2d*

Ejemplo: *En el experimento de Galton se estudia cómo afecta el tamaño de una semilla en el tamaño de sus descendientes. Se tienen los datos (diámetro de la semilla en mm)*

tamaño padres	15	16	17	18	19	20	21
tamaño hijos	15,4	15,7	16	16,3	16,6	17	17,3

Calcular las medias, dt y covarianza de las variables. Calcular la recta de regresión, y dibujar la gráfica correspondiente.

Solución: comenzamos cargando el paquete *stats* (si fuera necesario)

```
(%i1)  load(stats); fpprintprec:5;
(%i2)  x:[15,16,17,18,19,20,21]; y:[15.4,15.7,16,16.3,16.6,17,17.3]
(%i3)  z:transpose(matrix(x,y))
(%i4)  mean(z);   std(z)
```

$[18, 16.329]$ $\quad [2, \ 0.636]$

Es decir, $\bar{x} = 18$, $\sigma_x = 2$, $\quad \bar{y} = 16{,}329$, $\sigma_y = 0{,}636$

```
(%i5)   cov(z);        cor(z)
```

$$\begin{bmatrix} 4 & 1{,}2714 \\ 1{,}2714 & 0{,}405 \end{bmatrix} \qquad \begin{bmatrix} 1 & 0{,}999 \\ 0{,}999 & 1 \end{bmatrix}$$

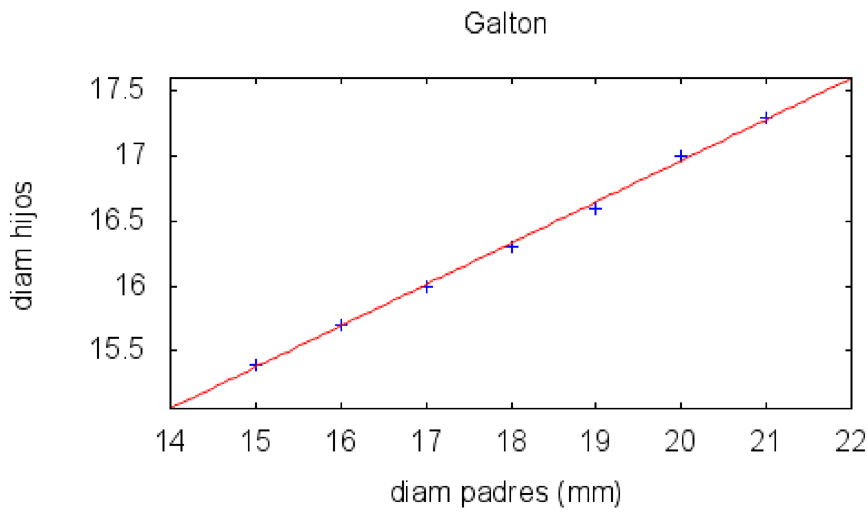
Por tanto concluimos $cov_{x,y} = 1{,}2714$ y $r = 0{,}999$ (muy buen ajuste).

```
(%i6)  simple_linear_regression(z)
```

model $= 0{,}318\,x + 10{,}607$	correlation $= 0{,}999$
$v_{\text{est}} = 0{,}00107$	$b_{\text{conf int}} = [0{,}302, 0{,}334]$
hypoth $H_0 : b = 0,\ H_1 : b \neq 0$	statistic=51.384
distribution $=$ [t–Stud, 5]	$p_{\text{val}} = 5{,}27711 \cdot 10^{-8}$

Es decir, la recta de regresión es $Y = 0{,}318X + 10{,}607$ y $r = 0'999$ (el resto de información no la usamos en este curso, pero la explico abajo)

```
(%i7)  wxdraw2d(points(z),color=red, explicit(0.318*X+10.607, X, 14,22),
       title="exper Galton",xlabel="diam padres (mm)",ylabel="diam hijos")
```



NOTA: los otros comandos que devuelve `simple_linear_regression` son

- $v_{\text{est}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = \text{ECM}_{n-2}$, es un estimador insesgado del ECM
- $b_{\text{conf int}} =$ intervalo de confianza (al 95 %) para la pendiente b
- hypoth H_0... test de hipótesis para $b = 0$; cuando $p_{\text{val}} < 0'001$ se rechaza la hipótesis $b = 0$, y se concluye con probabilidad alta que $b \neq 0$ (y por tanto, que existe dependencia lineal entre x e y).