



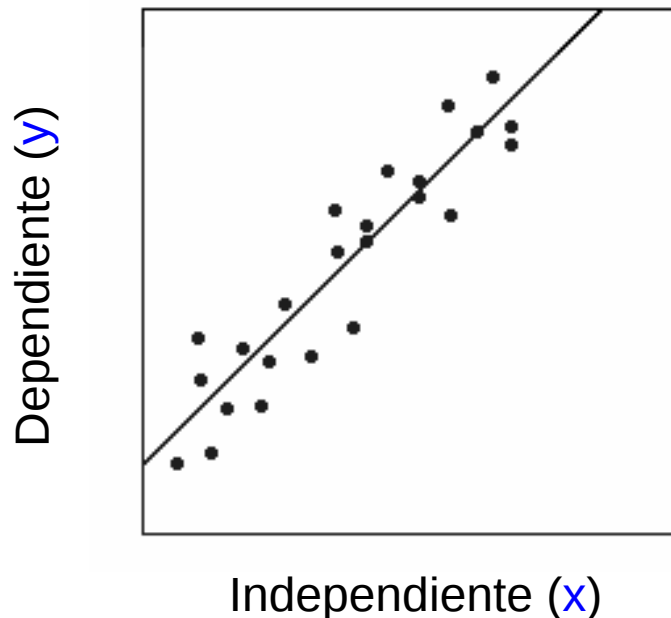
7.1. La ecuación lineal de regresión:

- Variable dependiente e independiente (fijas ó aleatorias):
 - . Fijas (modelo I de regresión)
 - . Aleatorias (modelo II; más complejo)
- Objetivo predictivo (básico en ciencia)
- Ecuación de regresión (mínimos cuadrados):
 - . Valor de intersección (a)
 - . Pendiente (b)

$$a = y - bx = \frac{\sum y_i - b \sum x_i}{N}$$

$$y = a + bx$$

$$b_{xy} = \frac{COV_{xy}}{\sigma^2 x} = \frac{\left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \right)}{\left(\frac{\sum (x_i - \bar{x})^2}{(n-1)} \right)}$$



Fórmulas simplificadas

$$Y = a + bX$$

$$b = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \quad a = \frac{\sum Y - b \sum X}{N}$$

Where,

N = number of observations, or years

X = a year index (decade)

Y = population size for given census years



7.1. La ecuación lineal de regresión:

- Aplicación a un ejemplo:
 - Gráfico dispersión
 - Estimación de recta predictiva

Porcentage de HDL en sangre (del mismo individuo)

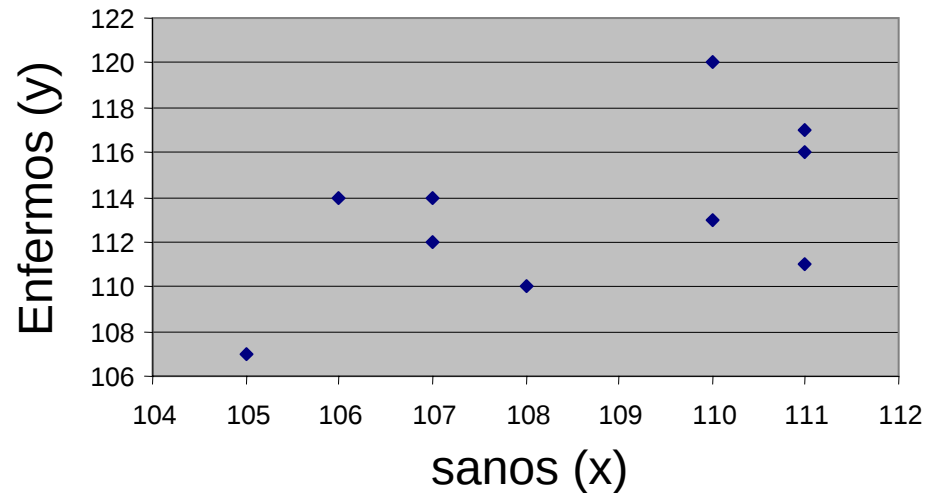
Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$Y = a + bX$ **Fórmulas simplificadas**

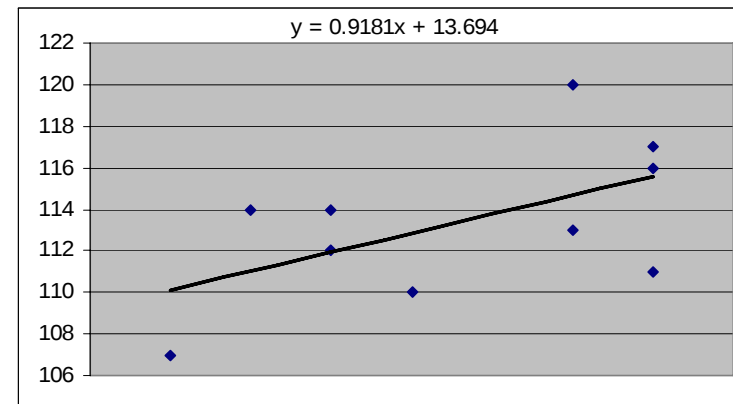
$$b = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \quad a = \frac{\sum Y - b \sum X}{N}$$

$$b = \text{cov}_{xy} / \text{Var}_x$$



$$\begin{aligned} \text{Cov}_{xy} &= 4,26 \\ V_x &= 4,64 \\ b_{xy} &= 0,918 \end{aligned}$$

$$\begin{aligned} \sum x_i &= 1086 \\ \sum y_i &= 1134 \\ a &= 13,69 \end{aligned}$$

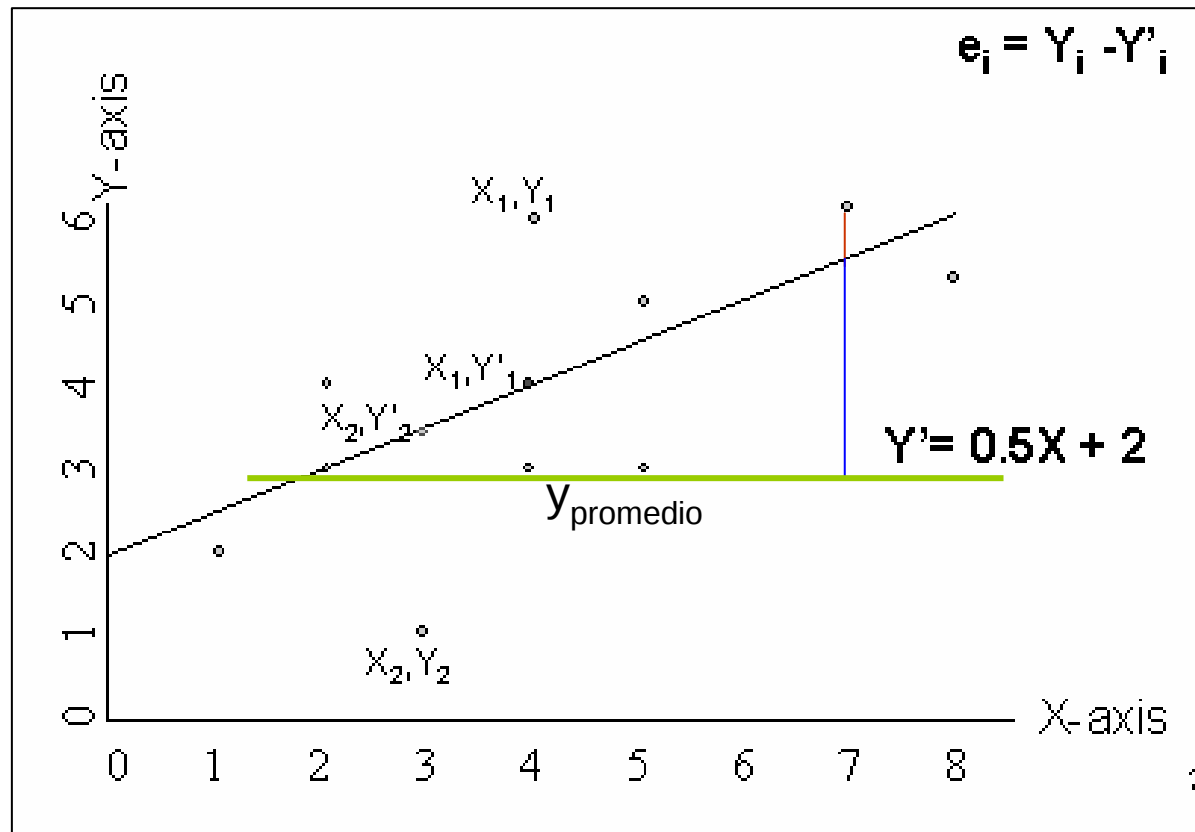




7.1. La ecuación lineal de regresión:

- Los residuos:

- . El valor predicho de y por la ecuación (y')
- . Porción de y explicado (azul) y no explicado (rojo)
- . Resíduo (e_i)



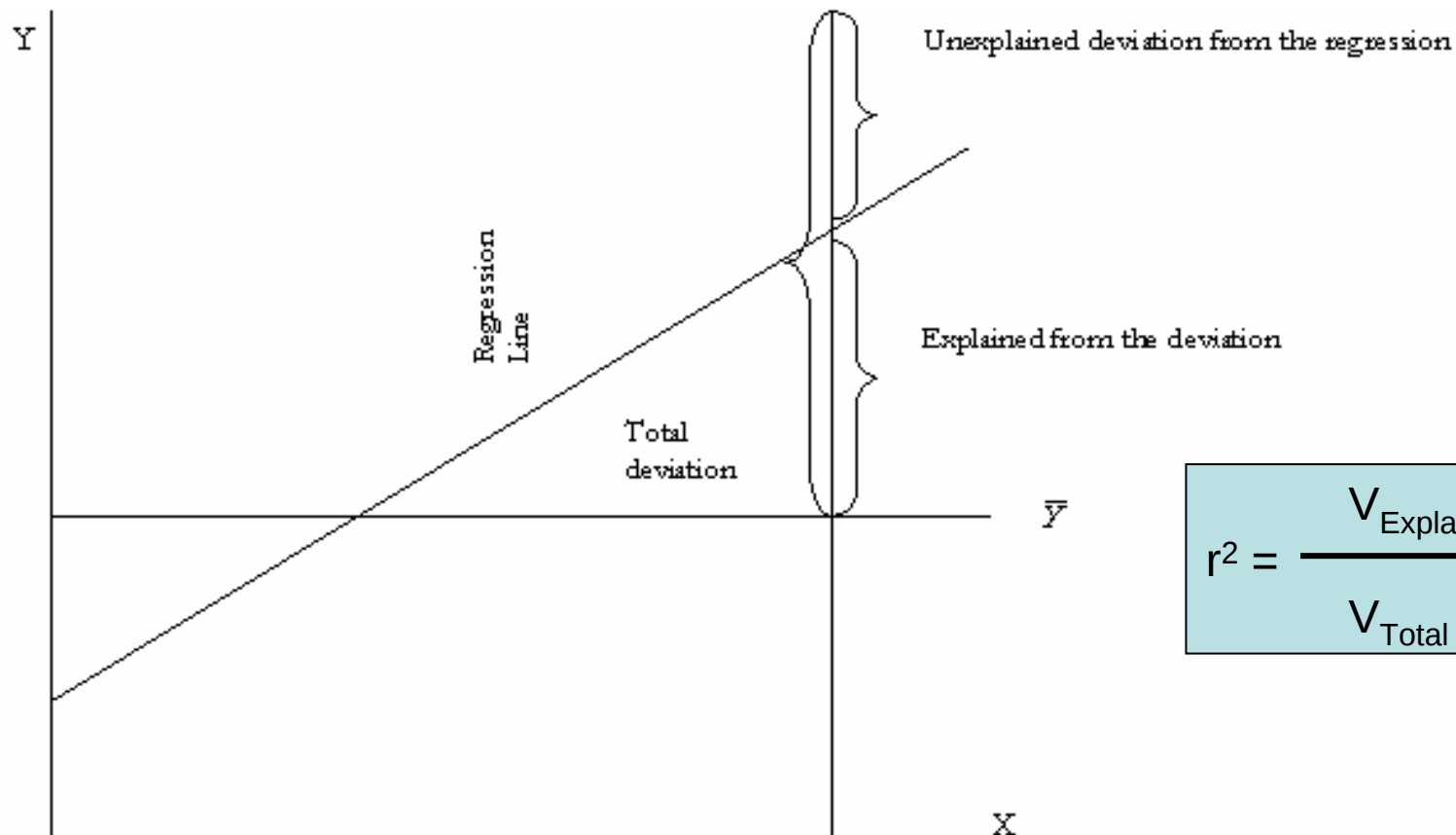


7.1. La ecuación lineal de regresión:

- La relación con el coeficiente de correlación:
 - . La r^2 (porcentaje de variación explicado)

$$r_{xy} = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

$$b_{xy} = r_{xy} (\sigma_y / \sigma_x)$$



$$r^2 = \frac{V_{\text{Explained}}}{V_{\text{Total}}}$$



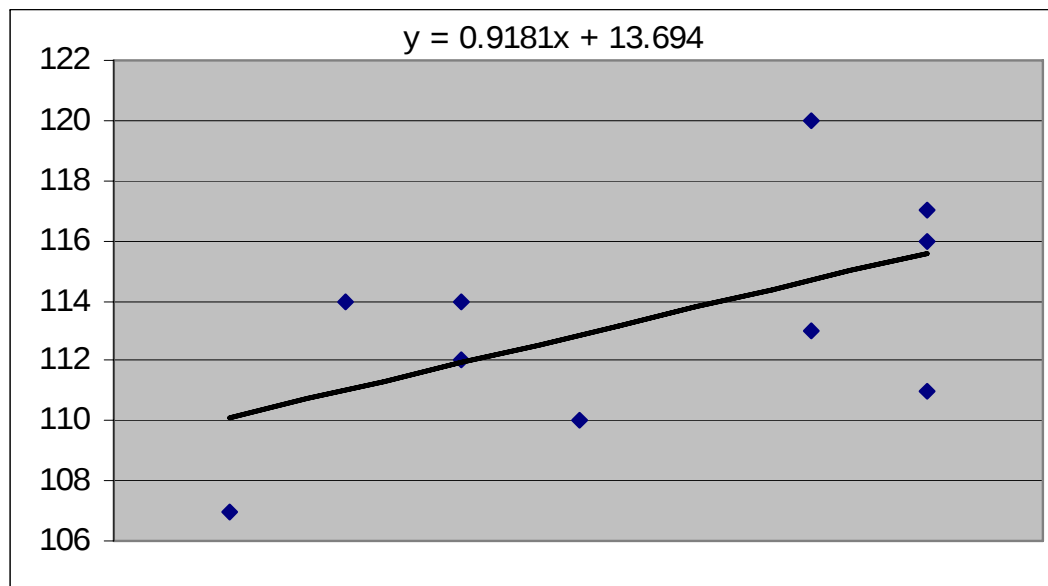
7.1. La ecuación lineal de regresión:

- La relación con el coeficiente de correlación:
 - . Ejemplo

Porcentaje de HDL en sangre (del mismo individuo)

Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107



$$r_{XY} = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

$$r = 0,56$$

$$r^2 = 0,31$$

% Explicado = 31%



7.2. El test estadístico asociado al análisis de regresión:

- El ANOVA de regresión (el mejor método de evaluación estadística):
 - . Partición variación explicada (Regresión) y No explicada (Error):
 - . SSr y Sse
 - . GLr y Gle
 - . Test F



7.2. El test estadístico asociado al análisis de regresión:

- El ANOVA de regresión:

. Ejemplo:

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$$SC_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SC_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	39.111	1	39.111	3.669	.092 ^a
	Residual	85.289	8	10.661		
	Total	124.400	9			

a. Variables predictoras: (Constante), sanos

b. Variable dependiente: enfermos

La ecuación de regresión **no explica una parte **significativa** de la variación**



7.2. El test estadístico asociado al análisis de regresión:

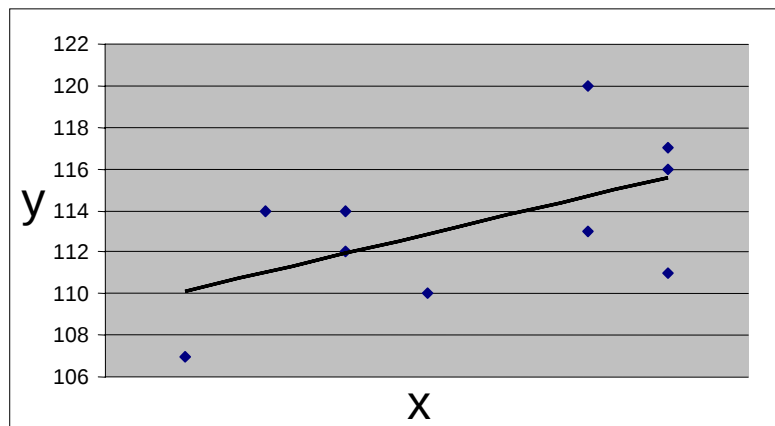
- Significación de b:
 - . Evaluación de σ_b
 - . Test t

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$$y = 13.694 + 0.9181x$$



$$\sigma_b^2 = \frac{CM_e}{\sum(x_i - \bar{x})^2}$$

$$\text{Test } t = \frac{b - 0}{\sigma_b}$$

$$CM_e = 10,66$$

$$\sum(x_i - \bar{x})^2 = 46,4$$

$$\sigma_b = \sqrt{(10,66/46,4)} = 0,479$$

$$t = 0,918/0,479 = 1,915$$

$$p = 0,092$$

La pendiente no se distingue significativamente de cero



7.2. El test estadístico asociado al análisis de regresión:

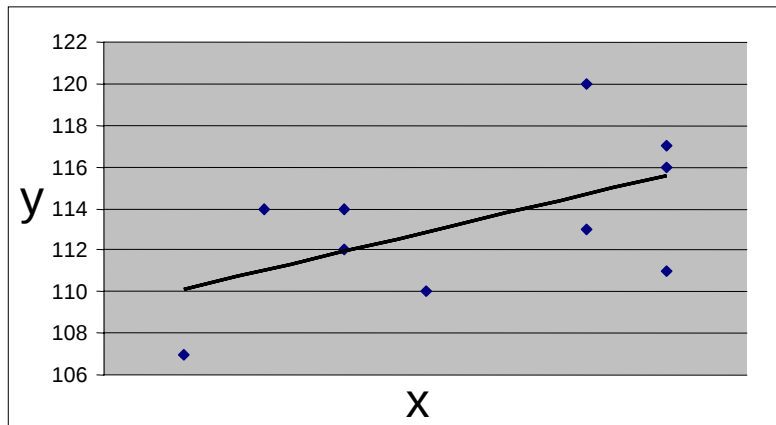
- Significación de a, o cualquier valor de y:
 - . Evaluación de σ_a
 - . Test t

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (t_1) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (t_2) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

$$y = 13.694 + 0.9181x$$



$$\sigma_a^2 = CM_e \left[\frac{1}{n} + \frac{(y-\bar{x})^2}{\sum x_i^2} \right]$$

$$\text{Test } t = \frac{a - 0}{\sigma_a}$$

$$\begin{aligned} CM_e &= 10,66 \\ y &= 0 \\ \bar{x}^2 &= 11793,96 \\ \sum x_i^2 &= 117986 \end{aligned}$$

$$\sigma_a = \sqrt{(10,66 \times (0,1 + 0,0999))} = 1,03$$

$$t = 13,694/1,03 = 13,29$$

La intersección en el eje y es significativamente distinta de cero



7.3. El uso práctico de los residuos de regresión:

- Uso práctico de los residuos:

- . Si tenemos más de dos variables explicativas (x y z independientes):
- . Se obtiene la regresión x/y
- . El residuo de la anterior se usa como dependiente para estudiar z:
 - . 1 (mucho) 2 (poco) 3 (nada)

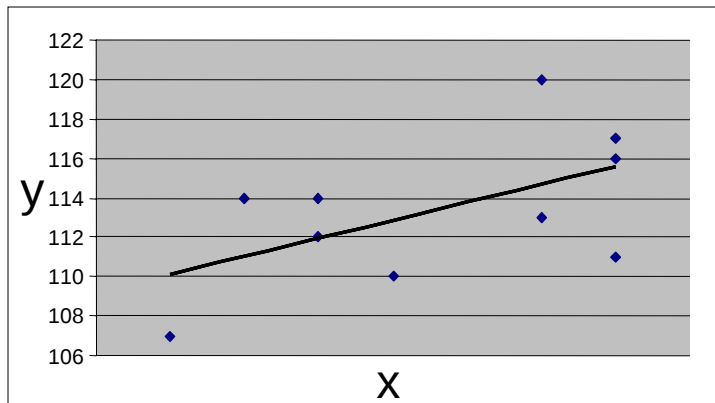
Porcentage de HDL en sangre (del mismo individuo)

Enfermos (y) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (x) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

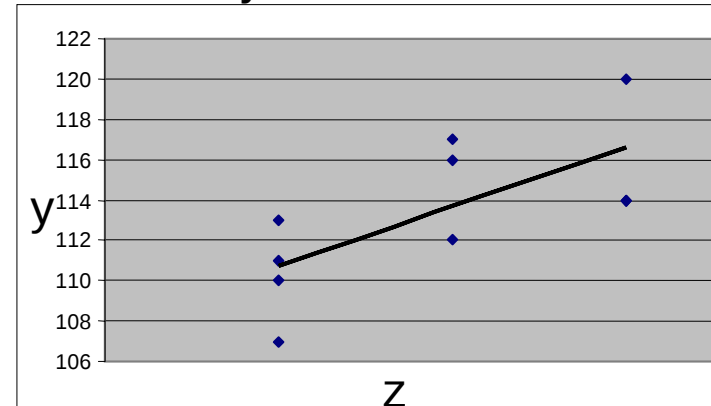
Fumador (z) : 3, 1, 2, 2, 3, 1, 1, 3, 3, 2

$$y = 13.694 + 0.9181x$$



$F_{xy} = 3,67^{ns}$
Explicado = 31%

$$y = 107.78 + 2.9565x$$



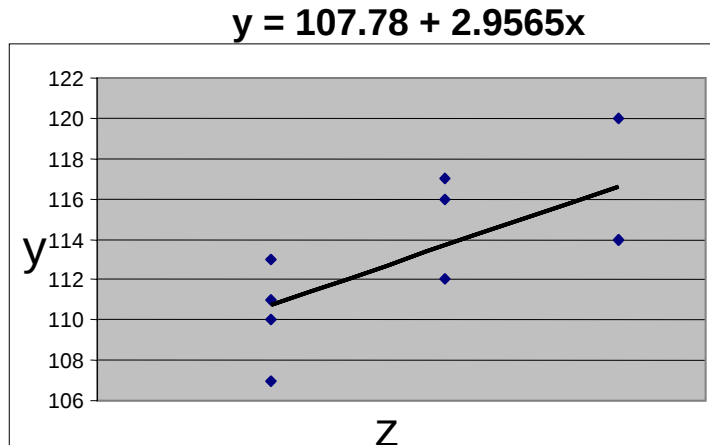
$F_{xz} = 7,53^*$
Explicado = 48%



7.3. El uso práctico de los residuos de regresión:

- Uso práctico de los residuos:

- . Si tenemos más de dos variables explicativas (x / z independientes):
- . Se obtiene la regresión x/y (fumar disminuye HDL)
- . El residuo de la anterior se usa como dependiente para estudiar z:



Porcentaje de HDL en sangre (del mismo individuo)

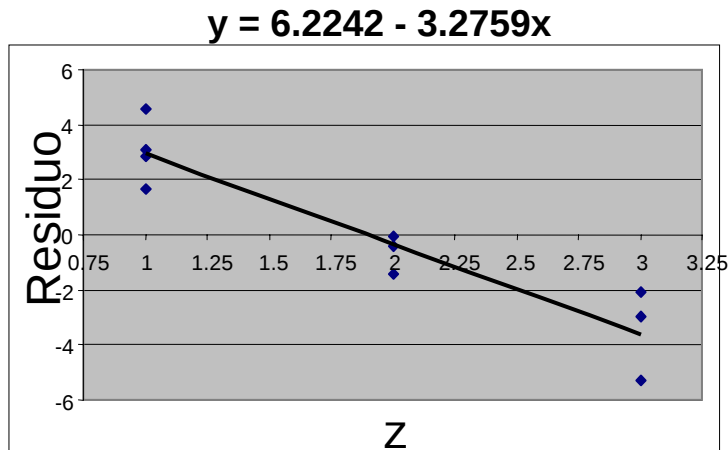
(y) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

(x) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

Res:

(z) : 3, 1, 2, 2, 3, 1, 1, 3, 3, 2

$F_{xy} = 52,7^{***}$
Explicado = 87%



Originalmente, parecía que cuanto menos se fumase más alto eran los valores de HDL

En la realidad si corregimos por variación en HDL: fumar aumenta realmente el valor de HDL



7.4. La regresión lineal multivariable:

- Objetivos:
 - . Conocer la capacidad predictiva de un conjunto de variables
 - . Inferir la mejor o mejores variables explicativas de un conjunto de variables
- Objetivo 1º: se introducen todas las variables:
 - . Extensión directa de la regresión con una variable independiente

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (y) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112

Sanos (x) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107

Fumador (z) : 3, 1, 2, 2, 3, 1, 1, 3, 3, 2

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	114.514	2	57.257	40.540	.000 ^a
	Residual	9.886	7	1.412		
	Total	124.400	9			

Explicada = 92%

		Coeficientes no estandarizados				
Modelo		B	Error típ.	sanos		
1	(Constante)	-11.383	19.259		-.591	.573
	fumador	3.336	.457	.786	7.307	.000
	sanos	1.091	.176	.666	6.195	.000



7.4. La regresión lineal multivariable:

- Objetivos:
 - . Conocer la capacidad predictiva de un conjunto de variables
 - . Inferir la mejor o mejores variables explicativas de un conjunto de variables
- Objetivo 2º: se usa el procedimiento paso a paso:
 - . Sólo se seleccionan algunas variables:
 - . Criterios de p de aceptación (0,05) y p de exclusión (0,10)

Porcentage de HDL en sangre (del mismo individuo)

Enfermos (y) : 120, 107, 110, 116, 114, 111, 113, 117, 114, 112
Sanos (x) : 110, 105, 108, 111, 107, 111, 110, 111, 106, 107
Fumador (z) : 3, 1, 2, 2, 3, 1, 1, 3, 3, 2

ANOVA^c

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	60.313	1	60.313	7.529	.025 ^a
	Residual	64.087	8	8.011		
	Total	124.400	9			
2	Regresión	114.514	2	57.257	40.540	.000 ^b
	Residual	9.886	7	1.412		
	Total	124.400	9			

a. Variables predictoras: (Constante), fumador

b. Variables predictoras: (Constante), fumador, sanos

c. Variable dependiente: enfermos

Explicada = 92%

Tema VII. La predicción de variables

7.4. La regresión lineal multivariable:

- Objetivos:
 - . Conocer la capacidad predictiva de un conjunto de variables
 - . Inferir la mejor o mejores variables explicativas de un conjunto de variables
- Objetivo 2º: se usa el procedimiento paso a paso:
 - . Sólo se seleccionan algunas variables:
 - . Criterios de p de aceptación (0,05) y p de exclusión (0,10)

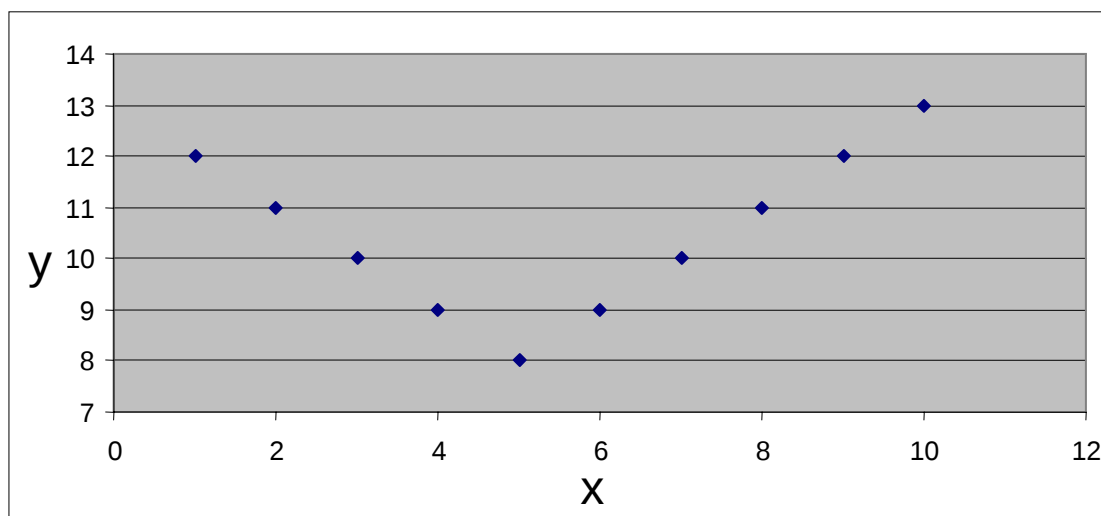


7.5. La regresión cuadrática:

- No todas las relaciones son lineales:
 - . Ecuaciones cuadráticas
 - . Idéntico procedimiento a las lineales

$$y = a + bx + cx^2$$

x	y
1	12
2	11
3	10
4	9
5	8
6	9
7	10
8	11
9	12
10	13



$F = 43,7^{***}$
Explicado = 92%

Coeficientes^a

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típ.	Beta		
1 (Constante)	13.833	.574		24.104	.000
x	-1.932	.240	-3.699	-8.060	.000
x2	.189	.021	4.093	8.919	.000

a. Variable dependiente: y

$$y = 13,83 - 1,9x + 0,2x^2$$



7.6. Las asunciones del modelo de regresión :

- Variables dependientes normales
- Variables independientes (fijas o normales)
- Un nº de casos muy superior al de variables (o regresión paso a paso)
- Cuando no se cumplan se puede:
 - . Aplicar algún test no paramétrico (método robusto de Kendall)
 - . Aplicar bootstrapping

Weight loss n mg (Y)	8.98	8.14	6.67	6.08	5.90	5.83	4.68	4.20	3.72
Percent relative humidity (X)	0.0	12.0	29.5	43.0	53.0	62.5	75.5	85.0	93.0

2. Estimate the Y-intercept: $Y_i - bX_i$ for $i = 1$ is 8.98000, for $i = 2$ is 8.79226, and so on

The median of the n intercepts is 8.78382.

$$y = 8,78 - 0,05x$$

The X-values are already arranged in ascending order of magnitude. If they had not been, then one should order them together with their paired Y-values.

Computation

1. Calculate $S_{ji} = (Y_j - Y_i)/(X_j - X_i)$ for all pairs X_i, X_j , where $j > i$.

$$S_{21} = (8.14 - 8.98)/(12.0 - 0.0) = -0.07000$$

$$S_{32} = (6.67 - 8.14)/(29.5 - 12.0) = -0.08400$$

.

.

.

$$S_{31} = (6.67 - 8.98)/(29.5 - 0.0) = -0.07831$$

.

.

.

$$S_{91} = (3.72 - 8.98)/(93.0 - 0.0) = -0.05656$$

The median of the $n(n-1)/2$ slopes is $b = -0.05436$.



Referencias Bibliográficas

LIBROS:

Siegel, S., Castellan, N.J. 1988. Nonparametric Statistics. McGraw Hill, New York

Sokal, R.R., Rohlf, F.J. 1995. Biometry. Freeman and co., New York

PÁGINAS WEB:

<http://www.cas.buffalo.edu/classes/psy/segal/2072001/regression/regression.html>

(Página Web donde se explica el análisis de regresión)