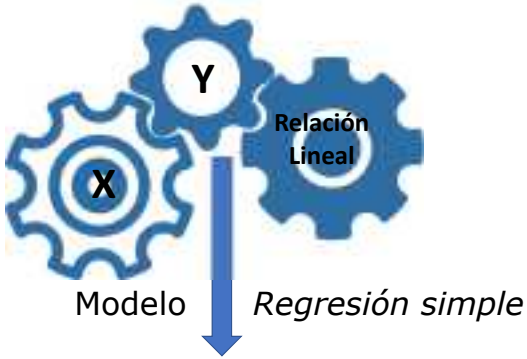
Tema 2

Descripción estadística de variables bidimensionales

Carlos Montes – uc3m

1. Introducción

Estudio de 2 caracteres simultáneos en cada elemento de la población.



2. Definiciones

- **Distribución conjunta de frecuencias de dos variables**

valores observados
y las frecuencias (relativas o absolutas)
de aparición de cada par.

Variable cualitativa ⇨ tabla de contingencia

$$\sum_i \sum_j fr(x_i, y_j) = 1$$

2. Definiciones

Distribución de frecuencias conjunta para las variables "número de hermanos" (columnas) y sexo (filas) de 95 estudiantes

Frequency Table for sexo by hermanos

	0	1	2	3	4	5	9	Row Total
0	3	13	11	2	2	0	1	32
	3,16%	13,68%	11,58%	2,11%	2,11%	0,00%	1,05%	33,68%
1	6	22	26	7	0	2	0	63
	6,32%	23,16%	27,37%	7,37%	0,00%	2,11%	0,00%	66,32%
Column Total	9	35	37	9	2	2	1	95
	9,47%	36,84%	38,95%	9,47%	2,11%	2,11%	1,05%	100,00%

Chicos

Chicas

2. Definiciones

- **Distribución marginal**

Distribución de cada una de las variables, consideradas por separado (distribución de los valores de una sin tener en cuenta los de la otra).

$$f(x_i) = \sum_j f(x_i, x_j)$$

$$f(y_j) = \sum_i f(x_i, x_j)$$

Aparece en los márgenes de la tabla.

2. Definiciones

Frequency Table for sexo by hermanos

	0	1	2	3	4	5	9	Row Total
0	3	13	11	2	2	0	1	32
	3,16%	13,68%	11,58%	2,11%	2,11%	0,00%	1,05%	33,68%
1	6	22	26	7	0	2	0	63
	6,32%	23,16%	27,37%	7,37%	0,00%	2,11%	0,00%	66,32%
Column Total	9	35	37	9	2	2	1	95
	9,47%	36,84%	38,95%	9,47%	2,11%	2,11%	1,05%	100,00%

Alumnos con 2 hermanos

2. Definiciones

- **Distribución condicionada** *de y para* $x=x_i$

es la distribución que se obtiene imponiendo la condición $x = x_i$

$$f_r(y_j|x = x_i) = \frac{f(x_i, y_j)}{f(x_i)}$$

$f_r(y_j|x_i = 2)$

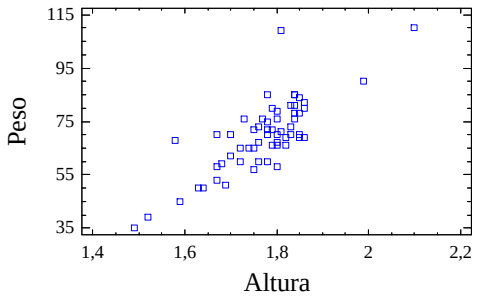
	0	1	2	3	4	5	9
0	3	13	11	2	2	0	1
1	6	22	26	7	0	2	0
	9	35	37	9	2	2	1

0	11/95 = 0,116
1	26/95 = 0,274
	0,39

0	11/37 = 0,298
1	26/37 = 0,702
	1

3. Representación gráfica

- *Diagrama de dispersión o nube de puntos*

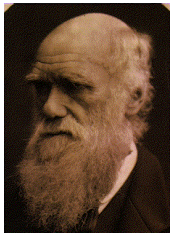


4.1. Covariacion. Tipos

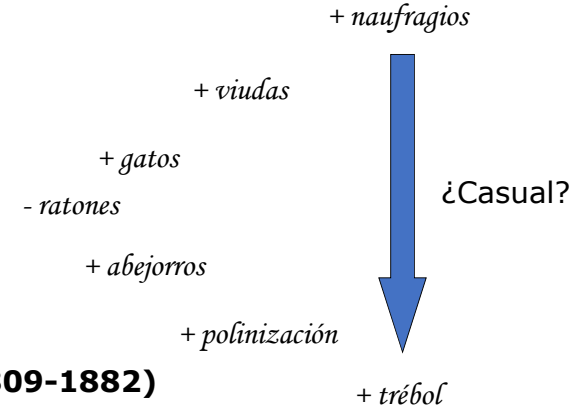
Variación conjunta o relación de dependencia entre las variables estudiadas (X, Y).

- Dependencia causal unilateral:
  X influye e Y, pero no la inversa.
- Interdependencia:
  X influye en Y, y viceversa.
- Dependencia indirecta:
  Las variables muestran una covariación a través de una tercera variable que influye en ellas.
- Concordancia.
- Covariación casual.

4.1. Covariacion. Tipos



Charles Darwin (1809-1882)

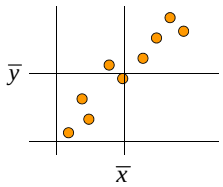


4.2. Covarianza

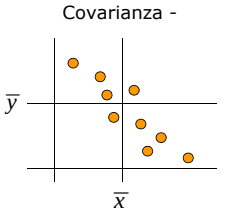
La *covarianza* es una medida descriptiva de la relación **lineal** entre cada par de variables.

$$\mathrm{cov}(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}$$

Covarianza +

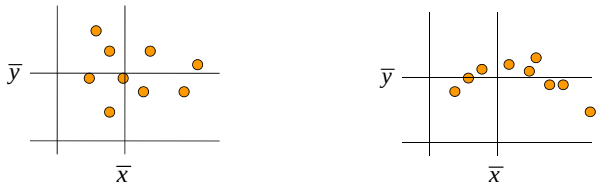


Covarianza -



4.2. Covarianza

Covarianza 0



4.2. Covarianza

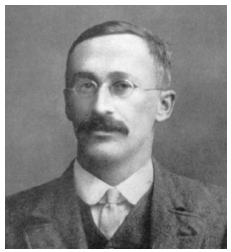
* La covarianza tiene el inconveniente de depender de las unidades de medida.

* Para evitarlo, se emplea el *coeficiente de correlación lineal r*.

4.3. Coeficiente de correlación

Sir Francis Galton (1822-1911)

$$r = \frac{S_{xy}}{S_x S_y}$$



Karl Pearson (1857-1936)

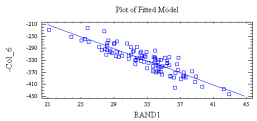
4.3. Coeficiente de correlación

r varía entre –1 y 1

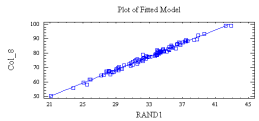
- *r*= -1 Correlación lineal perfecta e inversa. La nube de puntos es una recta de pendiente negativa.
- *r*= 1 Correlación lineal perfecta y directa. La nube de puntos es una recta de pendiente positiva.
- *r*= 0 No existe correlación, o bien existe una relación no lineal entre las variables.

4.3. Coeficiente de correlación

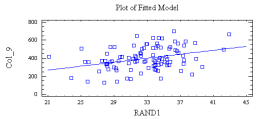
Algunos ejemplos numéricos:



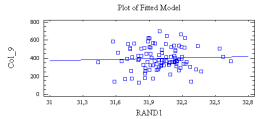
Correlation Coefficient = -0,889122



Correlation Coefficient = 0,994278



Correlation Coefficient = 0,340985



Correlation Coefficient = 0,0417867

4.4. Matriz de covarianzas

* Las medidas de dependencia lineal de un conjunto de datos bidimensionales pueden presentarse en forma de matriz.

$$M = \begin{pmatrix} s_x^{\,2} & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & s_y^{\,2} \end{pmatrix}$$

Matriz de covarianzas muestrales

$$R = \begin{pmatrix} 1 & corr(x,y) \\ corr(y,x) & 1 \end{pmatrix}$$

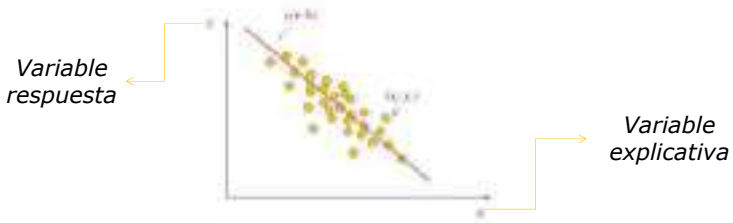
Matriz de correlaciones muestrales

corr(x,x)=corr(y,y)

5.1. Recta de regresión

* Recta que refleja, de la manera más aproximada posible, la evolución conjunta de dos variables.

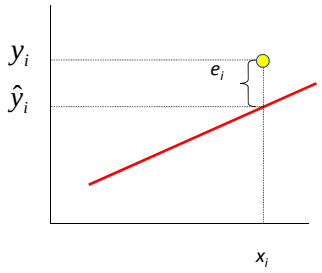
* Cuanto más próximo a ±1 esté el coeficiente de correlación, mayor será la capacidad de explicación de la recta.



5.1. Recta de regresión

Para cada x_i tendremos

- ordenada real y_i
- ordenada sobre la recta de regresión $\hat{y}_i$

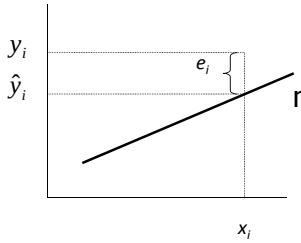


$$y_i = \hat{y}_i + e_i$$

$$e_i = y_i - \hat{y}_i$$

residuo

5.1. Recta de regresión



Posible criterio de construcción: minimizar la suma de los errores.

$$e_i = y_i - \hat{y}_i$$

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

Para evitar la influencia de los signos:

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

Método de los mínimos cuadrados
Carl Friedrich Gauss (1777-1855)

5.1. Recta de regresión

Si lo que queremos ajustar es una recta:

$$\hat{y}_i = a + bx_i$$

Minimizando llegamos a:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

(recta de regresión de Y sobre X)

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

(recta de regresión de X sobre Y)

5.2. Interpretación de los coeficientes

Como $y = a + bx$ $\quad \frac{dy}{dx} = b$

b es la pendiente de la recta:
incremento de y cuando *x* aumenta en una unidad.

$$\Delta\hat{y} = \hat{y}(x_i + 1) - \hat{y}(x_i) =$$
$$= [a + b(x_i + 1)] - [a + bx_i] = b$$

a es el valor de la recta cuando x=0

5.3. Evaluación del modelo

5.3. Evaluación del modelo

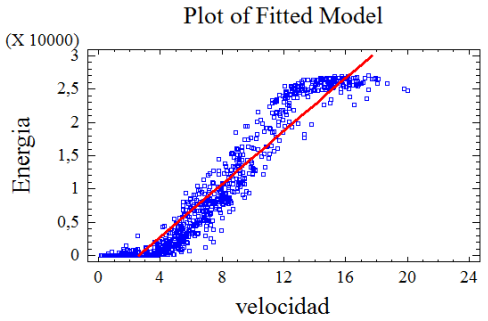
Gráfico de dispersión

Gráfico de valores previstos frente a valores observados

Gráfico de residuos frente a valores previstos

5.3. Evaluación del modelo

r= 0,96

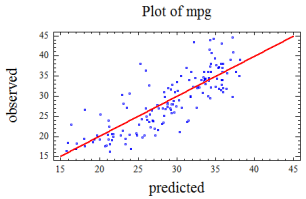


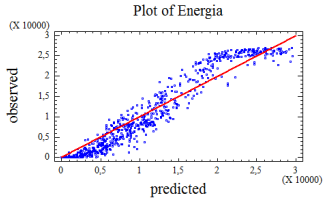
No hay relación lineal a pesar del elevado *r.*

5.3. Evaluación del modelo

Gráfico de valores previstos frente a valores observados

Linealidad ⇨ puntos distribuidos linealmente alrededor de la recta.

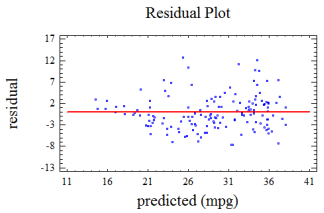




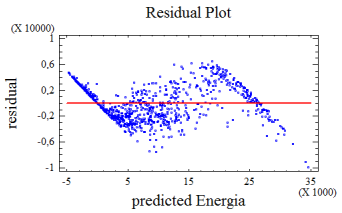
5.3. Evaluación del modelo

Gráfico de residuos frente a valores previstos

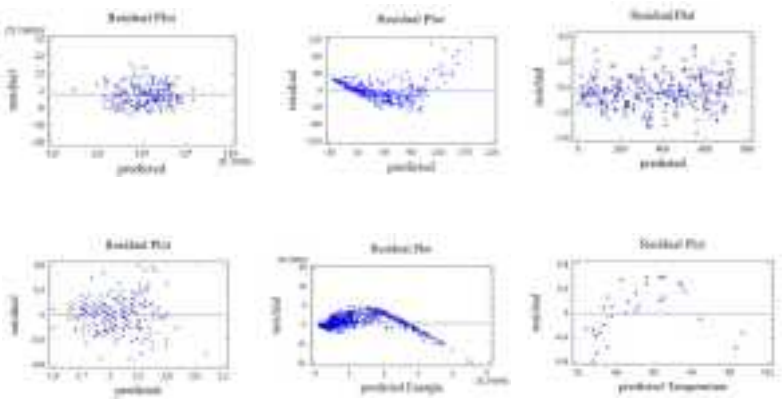
Linealidad ⇨ puntos distribuidos al azar



fuente: Weisberg, S (1985) Applied Linear Regression. John Wiley & Sons

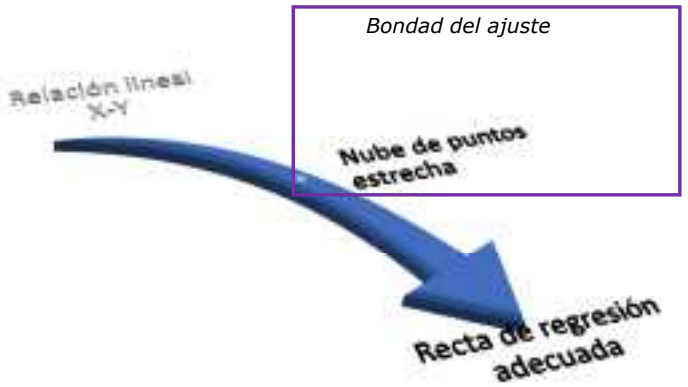


5.3. Evaluación del modelo



Carlos Montes – uc3m

5.4. Bondad del ajuste



5.4. Bondad del ajuste

* La regresión simple será tanto mejor cuanto más estrecha sea la nube de puntos alrededor de la muestra.

* La dispersión viene cuantificada por el coeficiente de correlación, o por el coeficiente de determinación R^2, que varía entre 0 y 1:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^{\,2}}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r^2$$

Variabilidad de los residuos

Variabilidad de los datos

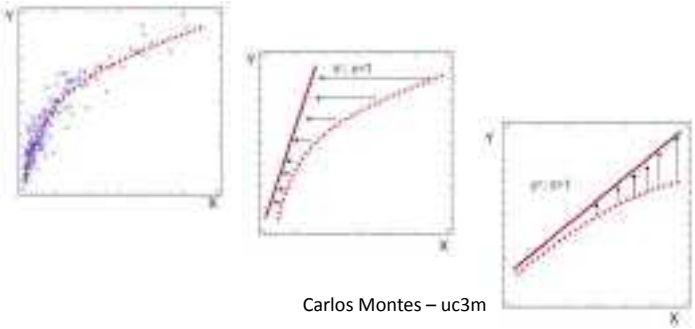
5.4. Bondad del ajuste

Cuanto más explicativa sea la regresión, menor será la variabilidad que queda en los residuos respecto a la de los datos, y R^2 será mayor.

Nos indica la proporción de la dispersión de la variable respuesta *y* que es capaz de explicar la recta de regresión.

6. Transformaciones

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis.



6. Transformaciones

Las más utilizadas son:

- Logaritmo

  $y = \ln x \quad x = \ln y$
- Potencia

  $y = x^c \quad x = y^c$
- Inversa

  $y = 1/x \quad x = 1/y$
- Raíz cuadrada

  $y = \sqrt{x} \quad x = \sqrt{y}$