BLOQUE 1
ESTADISTICA DESCRIPTIVA

OBJETIVOS:

EST. DESCRIPTIVA UNIVARIANTE

✓ Resumir y describir conjuntos de datos a través de distintos tipos de tablas, gráficos y medidas estadísticas.

✓ Estudiar relaciones entre las variables.
Realizar predicciones.

EST. DESCRIPTIVA BIVARIANTE

ESTADÍSTICA DESCRIPTIVA BIVARIANTE

1. Distribución Conjunta. Tabla de doble entrada
2. Diagrama de dispersión.
3. Medidas de Dependencia
4. Regresión Lineal.

- ¿Existe relacion entre el tiempo de conexión y la edad?
- ¿Cuánto tiempo estará conectada una persona de 25 años?
- ¿Y si es mujer?

SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.	SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.
Hombre	22	USA	2	76	Mujer	18	USA	6	58
Mujer	11	CHI	7	30	Hombre	16	USA	4	51
Hombre	18	CHI	6	55	Hombre	19	USA	3	39
Mujer	19	ESP	3	54	Mujer	12	ESP	2	33
Mujer	10	CHI	3	28	Mujer	21	ESP	7	56
Hombre	20	ESP	3	58	Hombre	20	ESP	4	83
Hombre	18	ESP	5	59	Mujer	18	ESP	5	63
Mujer	27	CHI	5	90	Mujer	24	USA	2	72
Hombre	15	USA	4	65	Mujer	17	ESP	2	67
Mujer	20	USA	5	55	Hombre	18	ESP	3	47
Mujer	18	ESP	2	57	Hombre	26	CHI	5	80
Hombre	20	ESP	3	54	Mujer	16	ESP	1	58
Mujer	24	ESP	4	77	Mujer	16	CHI	3	55
Hombre	17	USA	6	58	Hombre	18	USA	3	71
Hombre	23	USA	5	81	Hombre	16	ESP	4	57
Mujer	17	ESP	3	45	Hombre	20	USA	7	70
Hombre	20	USA	5	66	Hombre	16	ESP	1	57
Mujer	19	USA	6	61	Mujer	14	CHI	3	37
Hombre	21	CHI	3	61	Hombre	23	USA	5	78
Hombre	12	CHI	3	37	Mujer	24	USA	1	84
Mujer	23	ESP	3	60	Hombre	22	CHI	5	69
Mujer	21	ESP	4	69	Hombre	21	CHI	1	67
Hombre	19	ESP	4	78	Hombre	22	ESP	6	89
Mujer	23	USA	2	63	Hombre	17	CHI	7	61
Hombre	19	CHI	4	63	Hombre	25	ESP	2	88
Mujer	19	USA	3	54	Mujer	29	USA	4	80
Hombre	15	CHI	7	52	Hombre	23	ESP	7	83
Hombre	18	ESP	6	71	Hombre	18	CHI	6	51
Hombre	14	USA	7	41	Mujer	20	USA	5	51
Mujer	21	CHI	7	58	Mujer	21	CHI	2	49
Mujer	24	USA	3	70	Hombre	14	USA	4	46
Hombre	15	ESP	6	48	Mujer	17	USA	1	39
Mujer	18	CHI	4	63	Hombre	28	ESP	2	89
Mujer	21	ESP	4	56	Mujer	20	USA	5	66
Hombre	16	USA	7	46	Mujer	23	ESP	5	91
Hombre	11	ESP	2	48	Hombre	20	ESP	6	48
Hombre	18	USA	3	62	Hombre	19	CHI	4	57
Mujer	20	ESP	4	40	Mujer	19	CHI	1	51
Hombre	22	USA	3	54	Mujer	14	CHI	6	39

1.DISTRIBUCIÓN CONJUNTA

En ocasiones se hace necesario estudiar dos variables en el mismo conjunto de individuos. ¿De qué forma, desde el punto de vista descriptivo, podemos determinar si existe alguna relación entre ellas?

Tendremos un par de variables, (X, Y), tales que X tomará ***r*** valores distintos e Y tomará ***s*** valores distintos:

$$X \rightarrow x_1, x_2, \ldots, x_r$$

$$Y \rightarrow y_1, y_2, \ldots, y_s$$

En esta situación tendremos el par (x_i, y_j) un número determinado de veces n_{ij}, que llamaremos frecuencia absoluta del par.

¿Cuántos **hombres chinos** han visitado la Web? **11**

¿Cuántos usuarios de entre **10 y 15 años** han estado conectados de entre **25 y 35 minutos**? **3**

SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.	SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.
Hombre	22	USA	2	76	Mujer	18	USA	6	58
Mujer	11	CHI	7	30	Hombre	16	USA	4	51
Hombre	18	CHI	6	55	Hombre	19	USA	3	39
Mujer	19	ESP	3	54	Mujer	12	ESP	2	33
Mujer	10	CHI	3	28	Mujer	21	ESP	7	56
Hombre	20	ESP	3	58	Hombre	20	ESP	4	83
Hombre	18	ESP	5	59	Mujer	18	ESP	5	63
Mujer	27	CHI	5	90	Mujer	24	USA	2	72
Hombre	15	USA	4	65	Mujer	17	ESP	2	67
Mujer	20	USA	5	55	Hombre	18	ESP	3	47
Mujer	18	ESP	2	57	Hombre	26	CHI	5	80
Hombre	20	ESP	3	54	Mujer	16	ESP	1	58
Mujer	24	ESP	4	77	Mujer	16	CHI	3	55
Hombre	17	USA	6	58	Hombre	18	USA	3	71
Hombre	23	USA	5	81	Hombre	16	ESP	4	57
Mujer	17	ESP	3	45	Hombre	20	USA	7	70
Hombre	20	USA	5	66	Hombre	16	ESP	1	57
Mujer	19	USA	6	61	Mujer	14	CHI	3	37
Hombre	21	CHI	3	61	Hombre	23	USA	5	78
Hombre	12	CHI	3	37	Mujer	24	USA	1	84
Mujer	23	ESP	3	60	Hombre	22	CHI	5	69
Mujer	21	ESP	4	69	Hombre	21	CHI	1	67
Hombre	19	ESP	4	78	Hombre	22	ESP	6	89
Mujer	23	USA	2	63	Hombre	17	CHI	7	61
Hombre	19	CHI	4	63	Hombre	25	ESP	2	88
Mujer	19	USA	3	54	Mujer	29	USA	4	80
Hombre	15	CHI	7	52	Hombre	23	ESP	7	83
Hombre	18	ESP	6	71	Hombre	18	CHI	6	51
Hombre	14	USA	7	41	Mujer	20	USA	5	51
Mujer	21	CHI	7	58	Mujer	21	CHI	2	49
Mujer	24	USA	3	70	Hombre	14	USA	4	46
Hombre	15	ESP	6	48	Mujer	17	USA	1	39
Mujer	18	CHI	4	63	Hombre	28	ESP	2	89
Mujer	21	ESP	4	56	Mujer	20	USA	5	66
Hombre	16	USA	7	46	Mujer	23	ESP	5	91
Hombre	11	ESP	2	48	Hombre	20	ESP	6	48
Hombre	18	USA	3	62	Hombre	19	CHI	4	57
Mujer	20	ESP	4	40	Mujer	19	CHI	1	51
Hombre	22	USA	3	54	Mujer	14	CHI	6	39

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo. Distribución conjunta

PAÍS ORIG. SEXO	CHINA	ESPAÑA	USA	Tot. Fila
HOMBRE	11	16	15	**42**
MUJER	10	14	12	**36**
Tot. Col.	**21**	**30**	**27**	**78**

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	**3**
(35,45]	4	4	0	0	**8**
(45,55]	4	13	2	0	**19**
(55,65]	1	15	6	0	**22**
(65,75]	0	6	5	0	**11**
(75,85]	0	2	6	2	**10**
(85,95]	0	0	3	2	**5**
Tot. Col.	**12**	**40**	**22**	**4**	**78**

1.DISTRIBUCIÓN CONJUNTA

Si consideramos todas las frecuencias absolutas de todas las situaciones posibles obtenemos la ***distribución conjunta de (X,Y)***:

(X,Y)	y_1	...	y_j	...	y_s	
x_1	n_{11}	...	n_{1j}	...	n_{1s}	$n_{1\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
x_i	n_{i1}	...	n_{ij}	...	n_{is}	$n_{i\cdot}$
⋮	⋮	⋱	⋮	⋱	⋮	⋮
x_r	n_{r1}	...	n_{rj}	...	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot s}$	n

$$n_{i.} = \sum_{j=1}^{s} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}$$

$$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{s} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

Las distribuciones de frecuencias $n_{i\cdot}$, i=1,...,r y $n_{\cdot j}$, j=1,...,s se denominan **distribuciones marginales** de X e Y respectivamente. Concretamente, la primera y última columna de la tabla constituyen la ***distribución marginal de X***, y la primera y última fila la ***distribución marginal de Y***.

Escuela Superior de Ingeniería

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo. Distribuciones Marginales

PAÍS ORIG. SEXO	CHINA	ESPAÑA	USA	Tot. Fila
HOMBRE	11	16	15	**42**
MUJER	10	14	12	**36**
Tot. Col.	**21**	**30**	**27**	**78**

DISTRIB. MARGINALES

PAÍS	F.A.
CHI	21
ESP	30
USA	27
	78

SEXO	F.A.
HOMBRE	42
MUJER	36
	78

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	**3**
(35,45]	4	4	0	0	**8**
(45,55]	4	13	2	0	**19**
(55,65]	1	15	6	0	**22**
(65,75]	0	6	5	0	**11**
(75,85]	0	2	6	2	**10**
(85,95]	0	0	3	2	**5**
Tot. Col.	**12**	**40**	**22**	**4**	**78**

DISTRIB. MARGINALES

EDAD	F.A.
[10,15]	12
(15,20]	40
(20,25]	22
(25,30]	4
	78

T. CONEX.	F.A.
[25,35]	3
(35,45]	8
(45,55]	19
(55,65]	22
(65,75]	11
(75,85]	10
(85,95]	5
	78

1.DISTRIBUCIÓN CONJUNTA

Se denomina **frecuencia relativa** del par (x_i, y_j) a:

$$f_{ij} = \frac{n_{ij}}{n}$$

Y se verifica que: $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 1$

➢Ejemplo. Frecuencia relativa

PAÍS ORIG. SEXO	CHINA	ESPAÑA	ESTADOS UNIDOS	Tot. Fila
HOMBRE	11/78	16/78	15/78	**42**/78
MUJER	10/78	14/78	12/78	**36**/78
Tot. Col.	**21**/78	**30**/78	**27**/78	**78**/78

PAÍS ORIG. SEXO	CHINA	ESPAÑA	ESTADOS UNIDOS	Tot. Fila
HOMBRE	0,14	0,21	0,19	**0,54**
MUJER	0,13	0,18	0,15	**0,46**
Tot. Col.	**0,27**	**0,39**	**0,34**	**1,00**

1.DISTRIBUCIÓN CONJUNTA

En algunas ocasiones se posee información previa de una de las variables que puede modificar la información disponible de la otra. Es decir, buscamos la distribución de una variable en función de un valor fijo de la otra, a esto se le llama **distribución condicionada**.

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	**3**
(35,45]	4	4	0	0	**8**
(45,55]	4	13	2	0	**19**
(55,65]	1	15	6	0	**22**
(65,75]	0	6	5	0	**11**
(75,85]	0	2	6	2	**10**
(85,95]	0	0	3	2	**5**
Tot. Col.	**12**	**40**	**22**	**4**	**78**

¿Cómo se distribuye el tiempo de conexión entre los individuos con edades comprendidas entre 20 y 25 años?

¿Cómo se distribuye la edad entre los usuarios cuya sesión dura entre 45 y 55 minutos?

Las frecuencias condicionadas son:

$$f_{j|i} = \frac{n_{ij}}{n_{i.}} \qquad i = 1,\ldots,r \qquad\qquad f_{i|j} = \frac{n_{ij}}{n_{.j}} \qquad j = 1,\ldots,s$$

Escuela Superior de Ingeniería

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo. Distribución Condicionada

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	3
(35,45]	4	4	0	0	8
(45,55]	4	13	2	0	19
(55,65]	1	15	6	0	22
(65,75]	0	6	5	0	11
(75,85]	0	2	6	2	10
(85,95]	0	0	3	2	5
Tot. Col.	12	40	22	4	78

$X\|_{Y=(20,25]}$	
T. Conex./Edad=(20,25]	
T. Conex.	$f_{i\|j=3}$
[25,35]	0/22
(35,45]	0/22
(45,55]	2/22
(55,65]	6/22
(65,75]	5/22
(75,85]	6/22
(85,95]	3/22
Tot. Col.	22/22

$Y\|_{X=(45,55]}$	
Edad/ T. Conex=(45,55]	
Edad	$f_{j\|i=3}$
[10,15]	4/19
(15,20]	13/19
(20,25]	2/19
(25,30]	0/19
Tot. Col.	19/19

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo.

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	**3**
(35,45]	4	4	0	0	**8**
(45,55]	4	13	2	0	**19**
(55,65]	1	15	6	0	**22**
(65,75]	0	6	5	0	**11**
(75,85]	0	2	6	2	**10**
(85,95]	0	0	3	2	**5**
Tot. Col.	**12**	**40**	**22**	**4**	**78**

¿Cuántos usuarios de entre 15 y 20 años se conectan alrededor 1 hora?

¿Cuál es el porcentaje de usuarios que se conectan entre 35 y 45 minutos y tienen menos de 15 años?

¿Cuántos usuarios pasan conectados mas de 85 minutos?

¿Qué porcentaje de usuarios tienen una edad de entre 20 y 25 años?

¿Qué grupo de edad es el que más se conecta?

¿Qué porcentaje de usuarios que tienen una edad de entre 15 y 20 años pasa conectado entre 65 y 75 minutos?

Sabiendo que el tiempo de conexión de está entre 45 y 55 minutos ¿qué porcentaje de usuarios tienen entre 10 y 15 años?

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo.

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	3
(35,45]	4	4	0	0	8
(45,55]	4	13	2	0	19
(55,65]	1	15	6	0	22
(65,75]	0	6	5	0	11
(75,85]	0	2	6	2	10
(85,95]	0	0	3	2	5
Tot. Col.	12	40	22	4	78

¿Cuántos usuarios de entre 15 y 20 años se conectan alrededor 1 hora? n_{42}=15 usuarios.

¿Cuál es el porcentaje de usuarios que se conectan entre 35 y 45 minutos y tienen menos de 15 años? f_{21}=4/78=0,05=5%

¿Cuántos usuarios pasan conectados mas de 85 minutos? $n_{7.}$=5 usuarios.

¿Qué porcentaje de usuarios tienen una edad de entre 20 y 25 años? $f_{.3}$=22/78=0,28=28%

1.DISTRIBUCIÓN CONJUNTA

➢Ejemplo.

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	3	0	0	0	3
(35,45]	4	4	0	0	8
(45,55]	4	13	2	0	19
(55,65]	1	15	6	0	22
(65,75]	0	6	5	0	11
(75,85]	0	2	6	2	10
(85,95]	0	0	3	2	5
Tot. Col.	12	40	22	4	78

¿Qué grupo de edad es el que más se conecta? Los que tienen entre 15 y 20 años, 40 usuarios (MODA)

¿Qué porcentaje de usuarios que tienen una edad de entre 15 y 20 años pasa conectado entre 65 y 75 minutos?

$$f_{5|j=2}=6/40=0{,}15=15\%$$

Sabiendo que el tiempo de conexión de está entre 45 y 55 minutos ¿qué porcentaje de usuarios tienen entre 10 y 15 años?

$$f_{1|i=3}=4/19=0{,}21=21\%$$

Escuela Superior de Ingeniería

1.DISTRIBUCIÓN CONJUNTA

Decimos que la variable **X es independiente de la variable Y** si:

$$f_{i|j} = f_{i\times} \quad \forall i = 1,\ldots,r \quad \forall j = 1,\ldots,s$$

Es decir, la frecuencia condicionada coincide con la marginal.

➢Ejemplo. Independencia

PAÍS ORIG. EDAD	CHINA	ESPAÑA	USA	Tot. Fila
[10,20]	14	20	18	**52**
(20,30]	7	10	9	**26**
Tot. Col.	**21**	**30**	**27**	**78**

PAÍS ORIG. EDAD	Edad/País= CHI	Edad/País= ESP	Edad/País= USA	$f_{i\times}$
[10,20]	14/21=2/3	20/30=2/3	18/27=2/3	**52/78**=2/3
(20,30]	7/21=1/3	10/30=1/3	9/27=1/3	**26/78**=1/3

¡SON INDEPENDIENTES!

La medida de dependencia o independencia establece la información que se tiene de una de las variables en función del conocimiento que hay de la otra.

1.DISTRIBUCIÓN CONJUNTA

Decimos que la variable **X es independiente de la variable Y** si:

$$f_{ij} = f_{i.} \times f_{.j} \quad \forall i = 1,\ldots,r \ \text{ y } \ \forall j = 1,\ldots,s$$

➢Ejemplo. Independencia

Edad T. Conex.	[10,15]	(15,20]	(20,25]	(25,30]	Tot. Fila
[25,35]	0,04	0	0	0	**0,04**
(35,45]	0,05	0,05	0	0	**0,10**
(45,55]	0,05	**0,17**	0,02	0	**0,24**
(55,65]	0,01	0,19	0,08	0	**0,28**
(65,75]	0	0,08	0,06	0	**0,14**
(75,85]	0	0,02	0,08	0,03	**0,13**
(85,95]	0	0	0,04	0,03	**0,07**
Tot. Col.	**0,15**	**0,51**	**0,28**	**0,06**	**1**

$$f_{32} = 0,17 \neq 0,12 = 0,51 \cdot 0,24 = f_{3.} \cdot f_{.2}$$

¡NO SON INDEPENDIENTES!

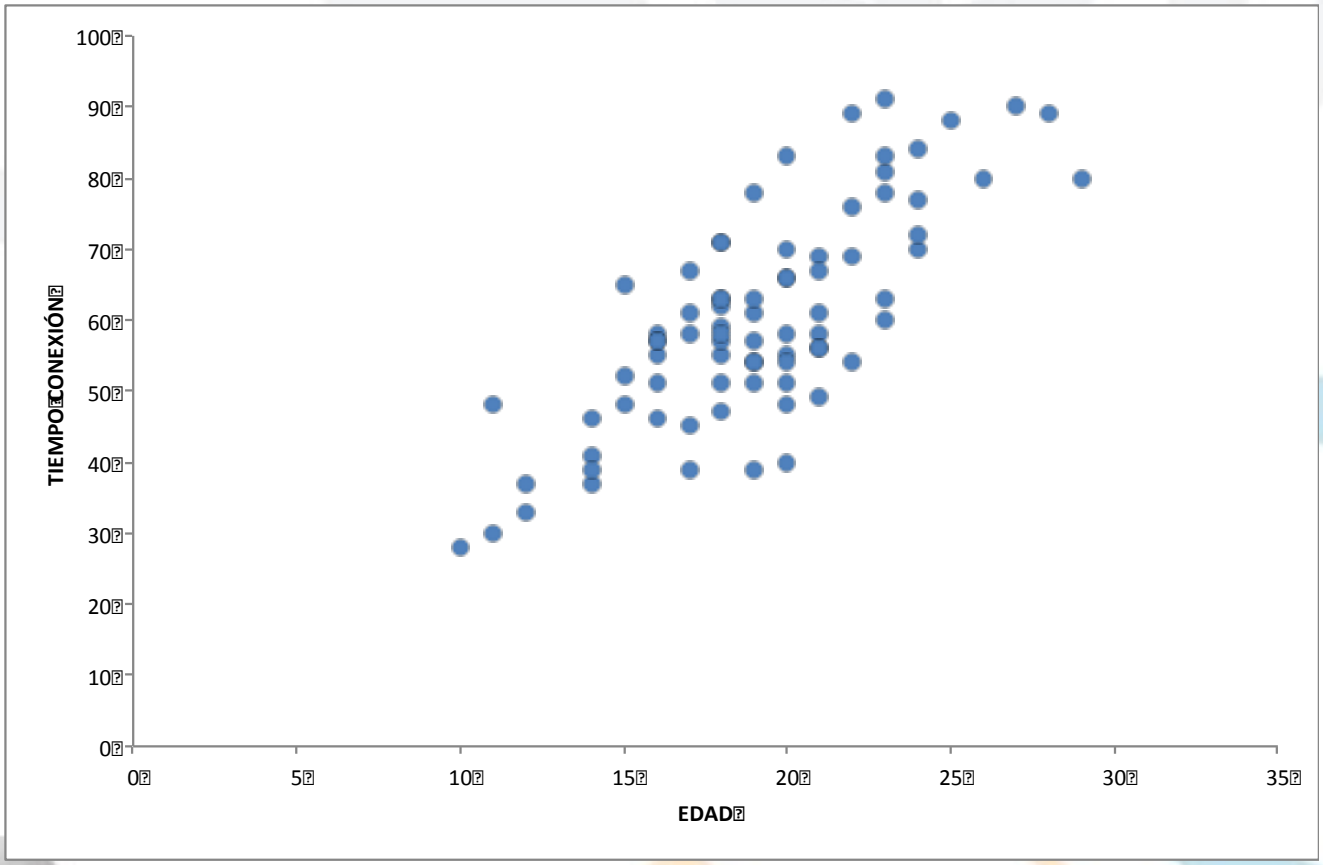
PAÍS ORIG. EDAD	Edad/País=CHI	Edad/País=ESP	Edad/País=USA	$f_{i\times}$
[10,20]	0,18	0,26	0,23	**0,67**
(20,30]	0,09	0,13	0,11	**0,33**
$f_{\times j}$	**0,27**	**0,39**	**0,34**	**1**

$$0,18 = f_{11} = f_{1\times} \times f_{\times 1} = 0,67 \times 0,27$$
$$0,26 = f_{12} = f_{1\times} \times f_{\times 2} = 0,67 \times 0,39$$
$$0,23 = f_{13} = f_{1\times} \times f_{\times 3} = 0,67 \times 0,34$$
$$0,09 = f_{21} = f_{2\times} \times f_{\times 1} = 0,33 \times 0,27$$
$$0,13 = f_{22} = f_{2\times} \times f_{\times 2} = 0,33 \times 0,39$$
$$0,11 = f_{23} = f_{2\times} \times f_{\times 3} = 0,33 \times 0,34$$

¡SON INDEPENDIENTES!

2.DIAGRAMA DE DISPERSIÓN

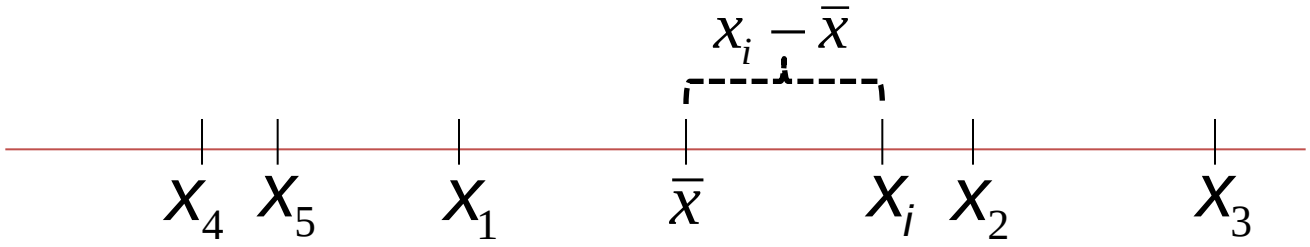
El diagrama de dispersión nos ayuda a ver cómo se distribuyen los datos y así poder ver si existe algún tipo de relación entre las características estudiadas, valores anómalos, etc.



EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

3.MEDIDAS DE DEPENDENCIA

Para el caso univariante se estudió una medida para expresar la variabilidad (dispersión) de los datos, la **varianza**.

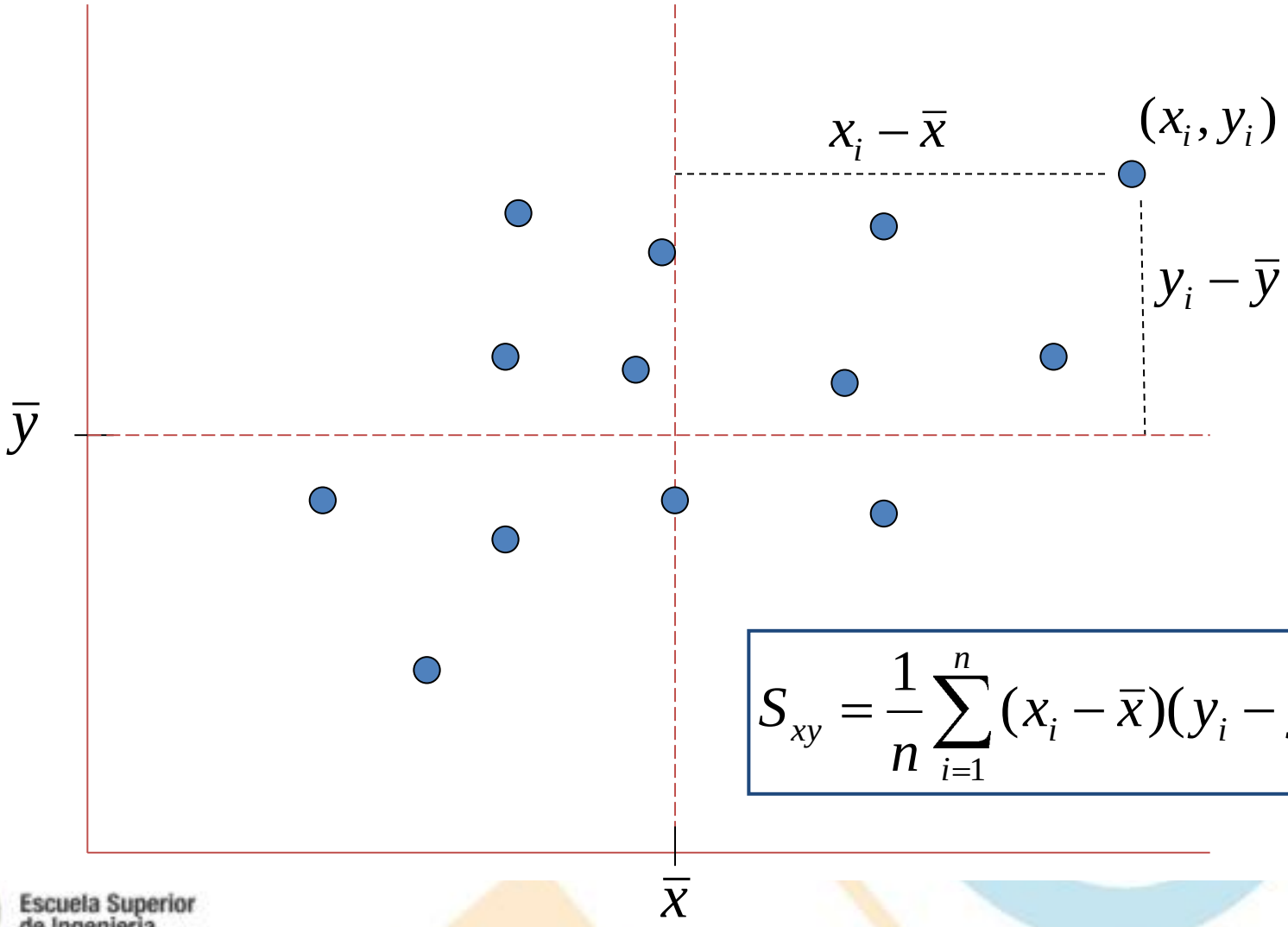


$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

¿Y para el caso bivariante?

3.MEDIDAS DE DEPENDENCIA

➢Covarianza



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

3.MEDIDAS DE DEPENDENCIA

➢Covarianza

La **covarianza, S_{XY}**, nos proporciona información sobre la variabilidad conjunta de dos variables numéricas.

Si las variables vienen dadas en una tabla de doble entrada su expresión es:

$$S_{XY} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})\, f_{ij}$$

Si no es así podemos expresarla como:

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

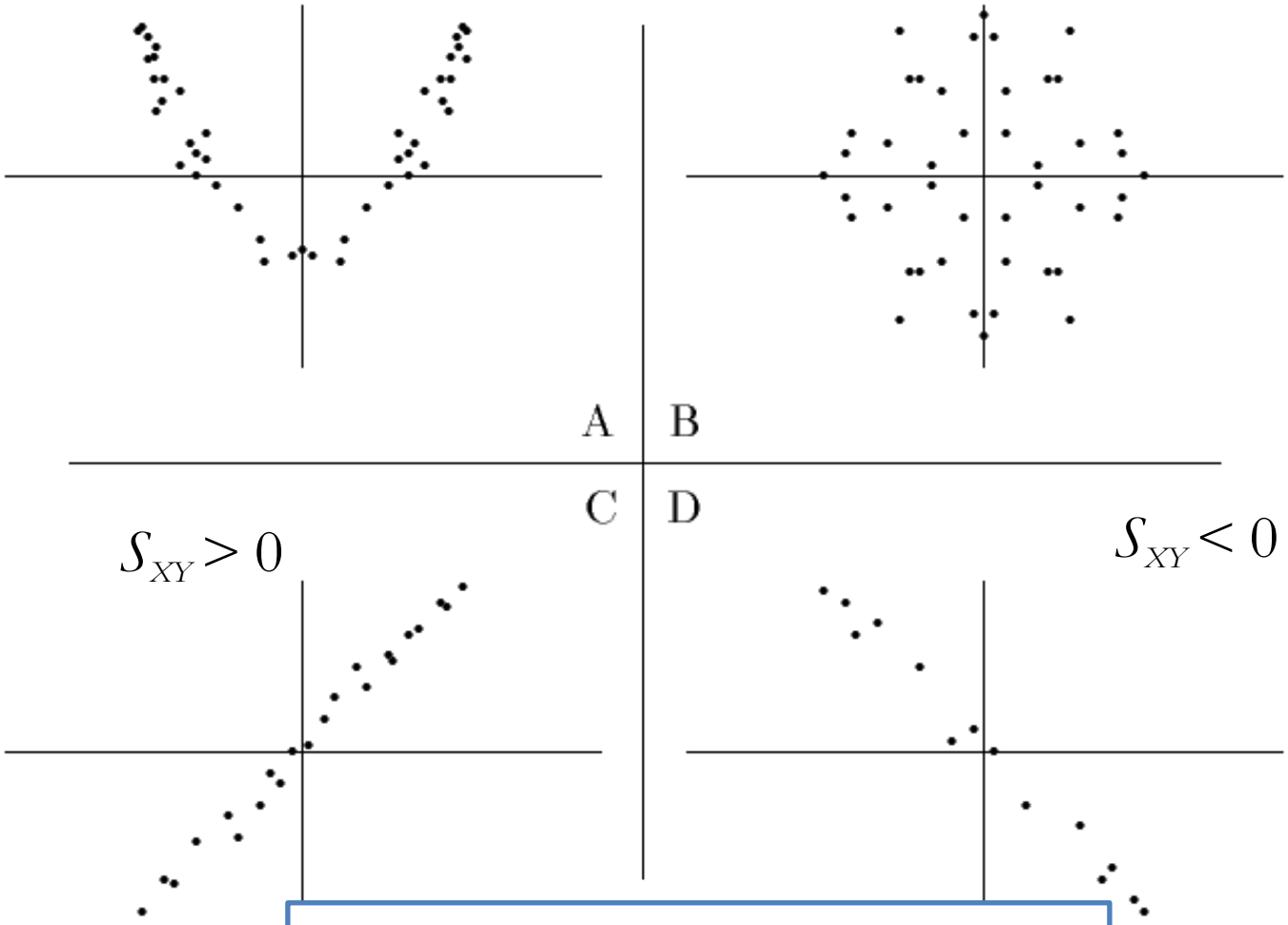
ó

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\cdot\bar{y}$$

- Si $S_{XY}>0$ es porque X e Y crecen o decrecen a la vez.
- Si $S_{XY}<0$ es porque cuando una de las dos variables crece, la otra tiende a decrecer.

3.MEDIDAS DE DEPENDENCIA

➢Covarianza



- Dependencia estadística lineal C y D
- Dependencia estadística parabólica A
- No existe dependencia estadística B

Escuela Superior de Ingeniería

3.MEDIDAS DE DEPENDENCIA

➢Ejemplo

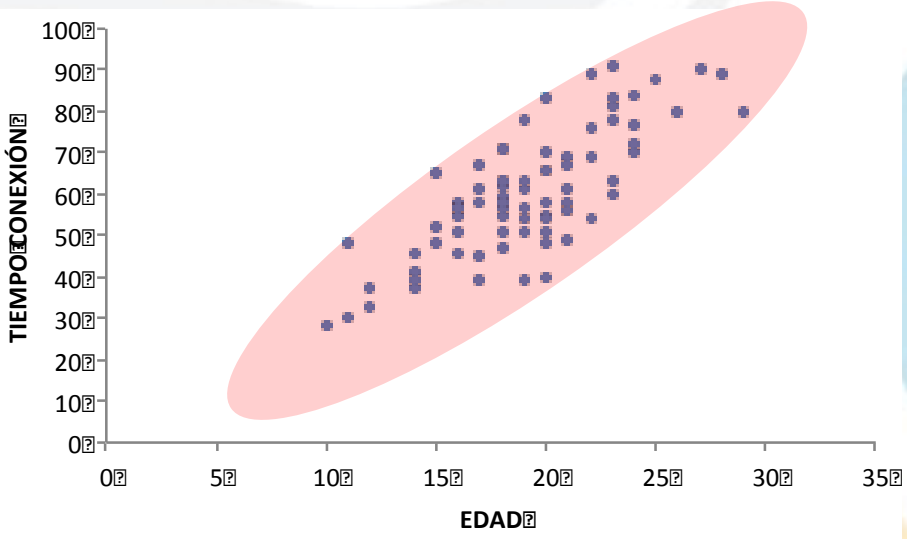
X → Edad
Y → Tiempo conexión

$$\bar{x} = 19{,}128$$

$$\bar{y} = 60{,}167$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$S_{xy} = \frac{1}{78}[22 \cdot 76 + 11 \cdot 30 + \cdots + 14 \cdot 39] - 19.128 - 60.127 = 44.459$$



EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

➢Coeficiente de Correlación de Pearson

La covarianza se ve afectada por las unidades en las que han sido medidas las variables.

Se define una medida adimensional que no se vea afectada por los cambios de unidad de medida:

Coeficiente de correlación de Pearson, r.

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

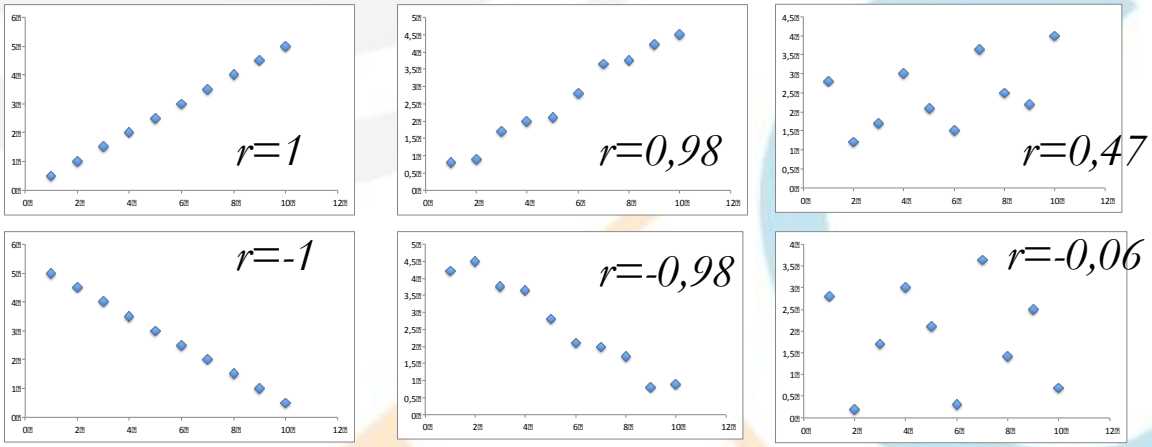
- Medida adimensional.
- Toma valores entre –1 y 1.
- Tiene el mismo signo que S_{XY}.

3.MEDIDAS DE DEPENDENCIA

➢Coeficiente de Correlación de Pearson

- Si r próximo a 1: ***relación lineal y directa*** entre las variables.
- Si r próximo a –1: ***relación lineal e inversa*** entre las variables.
- Si r próximo a 0: variables están ***incorreladas*.**

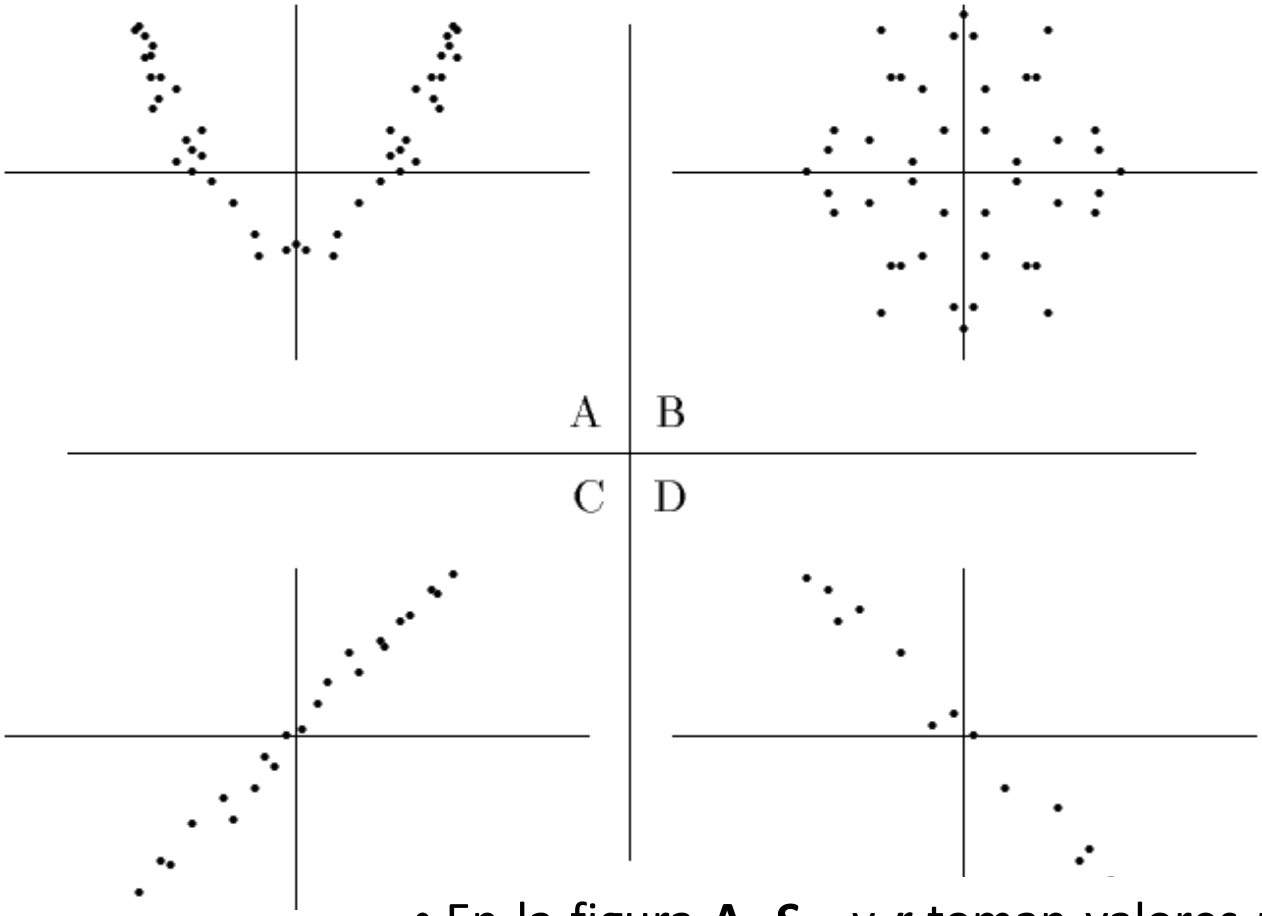
$$r = \frac{S_{xy}}{S_x S_y}$$



Escuela Superior de Ingeniería

3.MEDIDAS DE DEPENDENCIA

➢Coeficiente de Correlación de Pearson



En la figura **B**, no existe ningún tipo de relación estadística.

• En la figura **A**, S_{XY} y r toman valores próximos a 0, por tanto no existe relación lineal.

• Esto no descarta que pueda existir otro tipo de relación estadística, por ejemplo, parabólica.

3.MEDIDAS DE DEPENDENCIA

➢Ejemplo. Coeficiente de Correlación de Pearson

X → Edad

Y → Tiempo conexión

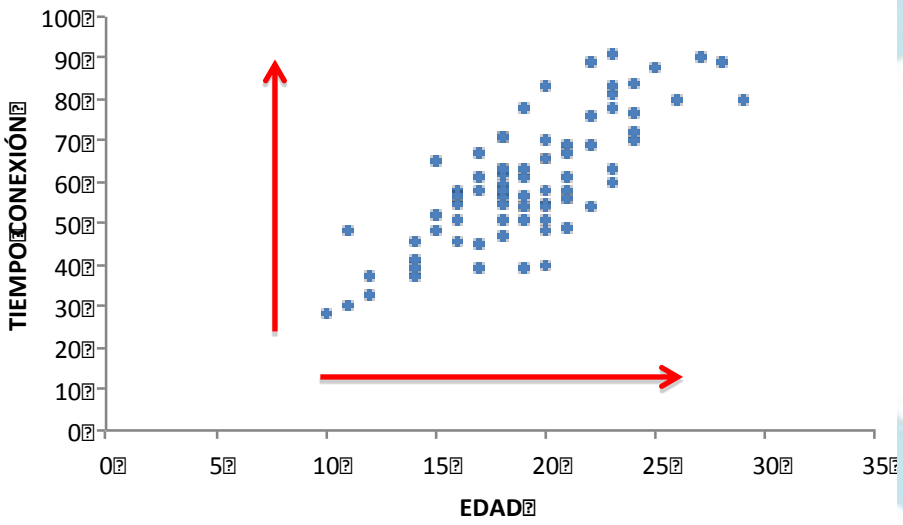
$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 3.846$$

$$\bar{x} = 19{,}128$$

$$\bar{y} = 60{,}167$$

$$S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 14.950$$

$$r = \frac{S_{XY}}{S_X \times S_y} = \frac{44{,}459}{3{,}846 \times 14{,}950} = 0{,}7733$$



EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

➢Coeficiente de correlación de Spearman

El **coeficiente de correlación por rangos de Spearman** es un coeficiente que se utiliza cuando queremos medir el grado de asociación entre dos variables que toman valores **ordinales**.

Su interpretación es análoga a la del coeficiente de correlación de Pearson ya que además toma valores entre –1 y 1.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Donde:

r_s : es el coeficiente de correlación por rangos de Spearman

d_i : es la diferencia entre el valor ordinal de la variable X y el de la variable Y en el elemento i-ésimo.

n : es el tamaño de la muestra.

➢Coeficiente de Contingencia (v. cualitativas)

El **coeficiente de contingencia, C,** mide el grado de asociación entre variables de tipo **cualitativo**.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

n es el tamaño de la muestra.

Su cálculo se basa en otro coeficiente, χ^2 , que está basado en la comparación de frecuencias observadas y esperadas al organizar los datos en una tabla de contingencia.

El coeficiente χ^2 no está acotado superiormente, sin embargo, el coeficiente de contingencia toma valores entre 0 y 1, indicando la intensidad de la relación pero no su sentido.

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$O_{ij} = n_{ij}$ son las frecuencias observadas

$E_{ij} = \frac{n_{i\times} \times n_{\times j}}{n}$ son las frecuencias esperadas o teóricas

➢Ejemplo. Coeficiente de Contingencia

Tabla de Contingencia

Frecuencias observadas

O_{ij}	CHINA	ESPAÑA	USA	Tot. Fila
HOMBRE	11	16	15	42
MUJER	10	14	12	36
Tot. Col.	21	30	27	78

$$E_{11} = \frac{21 \times 42}{78} = 11{,}31$$

Frecuencias esperadas

$E_{ij} = \frac{n_{i\times} \times n_{\times j}}{n}$	CHINA	ESPAÑA	ESTADOS UNIDOS
HOMBRE	11,31	16,15	14,54
MUJER	9,69	13,85	12,46

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(11-11.31)^2}{11.31} + \frac{(16-16.15)^2}{16.15} + \cdots + \frac{(12-12.46)^2}{12.46} = 0.05$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{0.05}{0.05+78}} = 0.03$$

En este caso, el valor del coeficiente indica poca asociación entre las variables *PAÍS* y *SEXO*.

Escuela Superior de Ingeniería

EJERCICIO 1

El complemento mensual de productividad (euros) que perciben los trabajadores de una empresa de creación de software según el número de horas trabajadas semanalmente se distribuye de acuerdo con la siguiente tabla:

Horas sem/complemento	70-110	110-150	150-170	170-190	190-250
31-35	5	4	2	1	0
35-37	1	2	4	3	3
37-41	0	3	4	2	6

(a) El número medio de horas trabajadas semanalmente. Si se hicieran dos horas extras en todas las empresas ¿cuál sería la nueva media?

(b) ¿Entre qué valores se encuentra el complemento más habitual para los empleados que trabajan al menos 35 horas semanalmente?.

(c) El complemento mediano de los empleados que trabajan al menos 35 horas.

EJERCICIO 1

El complemento mensual de productividad (euros) que perciben los trabajadores de una empresa de creación de software según el número de horas trabajadas semanalmente se distribuye de acuerdo con la siguiente tabla:

Horas sem/complemento	70-110	110-150	150-170	170-190	190-250
31-35	5	4	2	1	0
35-37	1	2	4	3	3
37-41	0	3	4	2	6

(d)La distribución del nº horas para complementos menores o iguales a 150 €.
(e) Consideramos en la empresa dos grupos de trabajadores: el grupo A, que incluye a los que trabajan como mucho 35 horas semanales, y el grupo B, que incluye al resto de trabajadores. ¿En cuál de los dos grupos es más equitativo el reparto de salarios?.

Sol. (a) 36;225 y 38;225, respectivamente; (b) 190-250 euros; (c) 170 euros; (d) horas/complemento <150 : 9=15; 3=15; 3=15; (e) En el grupo B.

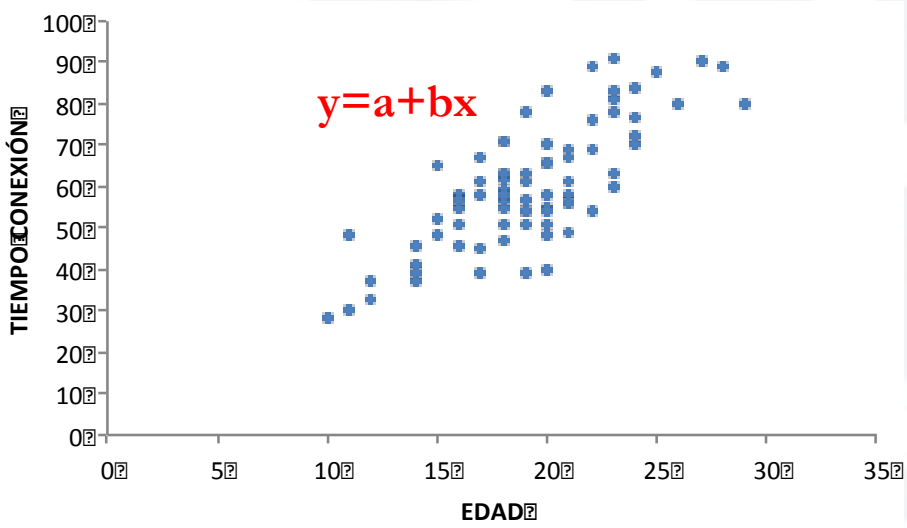
Escuela Superior de Ingeniería

- ¿Existe relacion entre el tiempo de conexión y la edad?
- ¿Cuánto tiempo estará conectada una persona de 25 años?

SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.	SEXO	EDAD	PAIS ORIGEN	Nº CONEX. SEMANALES	TIEMPO CONEX.
Hombre	22	USA	2	76	Mujer	18	USA	6	58
Mujer	11	CHI	7	30	Hombre	16	USA	4	51
Hombre	18	CHI	6	55	Hombre	19	USA	3	39
Mujer	19	ESP	3	54	Mujer	12	ESP	2	33
Mujer	10	CHI	3	28	Mujer	21	ESP	7	56
Hombre	20	ESP	3	58	Hombre	20	ESP	4	83
Hombre	18	ESP	5	59	Mujer	18	ESP	5	63
Mujer	27	CHI	5	90	Mujer	24	USA	2	72
Hombre	15	USA	4	65	Mujer	17	ESP	2	67
Mujer	20	USA	5	55	Hombre	18	ESP	3	47
Mujer	18	ESP	2	57	Hombre	26	CHI	5	80
Hombre	20	ESP	3	54	Mujer	16	ESP	1	58
Mujer	24	ESP	4	77	Mujer	16	CHI	3	55
Hombre	17	USA	6	58	Hombre	18	USA	3	71
Hombre	23	USA	5	81	Hombre	16	ESP	4	57
Mujer	17	ESP	3	45	Hombre	20	USA	7	70
Hombre	20	USA	5	66	Hombre	16	ESP	1	57
Mujer	19	USA	6	61	Mujer	14	CHI	3	37
Hombre	21	CHI	3	61	Hombre	23	USA	5	78
Hombre	12	CHI	3	37	Mujer	24	USA	1	84
Mujer	23	ESP	3	60	Hombre	22	CHI	5	69
Mujer	21	ESP	4	69	Hombre	21	CHI	1	67
Hombre	19	ESP	4	78	Hombre	22	ESP	6	89
Mujer	23	USA	2	63	Hombre	17	CHI	7	61
Hombre	19	CHI	4	63	Hombre	25	ESP	2	88
Mujer	19	USA	3	54	Mujer	29	USA	4	80
Hombre	15	CHI	7	52	Hombre	23	ESP	7	83
Hombre	18	ESP	6	71	Hombre	18	CHI	6	51
Hombre	14	USA	7	41	Mujer	20	USA	5	51
Mujer	21	CHI	7	58	Mujer	21	CHI	2	49
Mujer	24	USA	3	70	Hombre	14	USA	4	46
Hombre	15	ESP	6	48	Mujer	17	USA	1	39
Mujer	18	CHI	4	63	Hombre	28	ESP	2	89
Mujer	21	ESP	4	56	Mujer	20	USA	5	66
Hombre	16	USA	7	46	Mujer	23	ESP	5	91
Hombre	11	ESP	2	48	Hombre	20	ESP	6	48
Hombre	18	USA	3	62	Hombre	19	CHI	4	57
Mujer	20	ESP	4	40	Mujer	19	CHI	1	51
Hombre	22	USA	3	54	Mujer	14	CHI	6	39

4.REGRESIÓN LINEAL

Si tenemos dos variables cuantitativas, ¿existe alguna forma de establecer alguna relación entre ambas que permita expresar los valores de una de ellas en función de la otra?



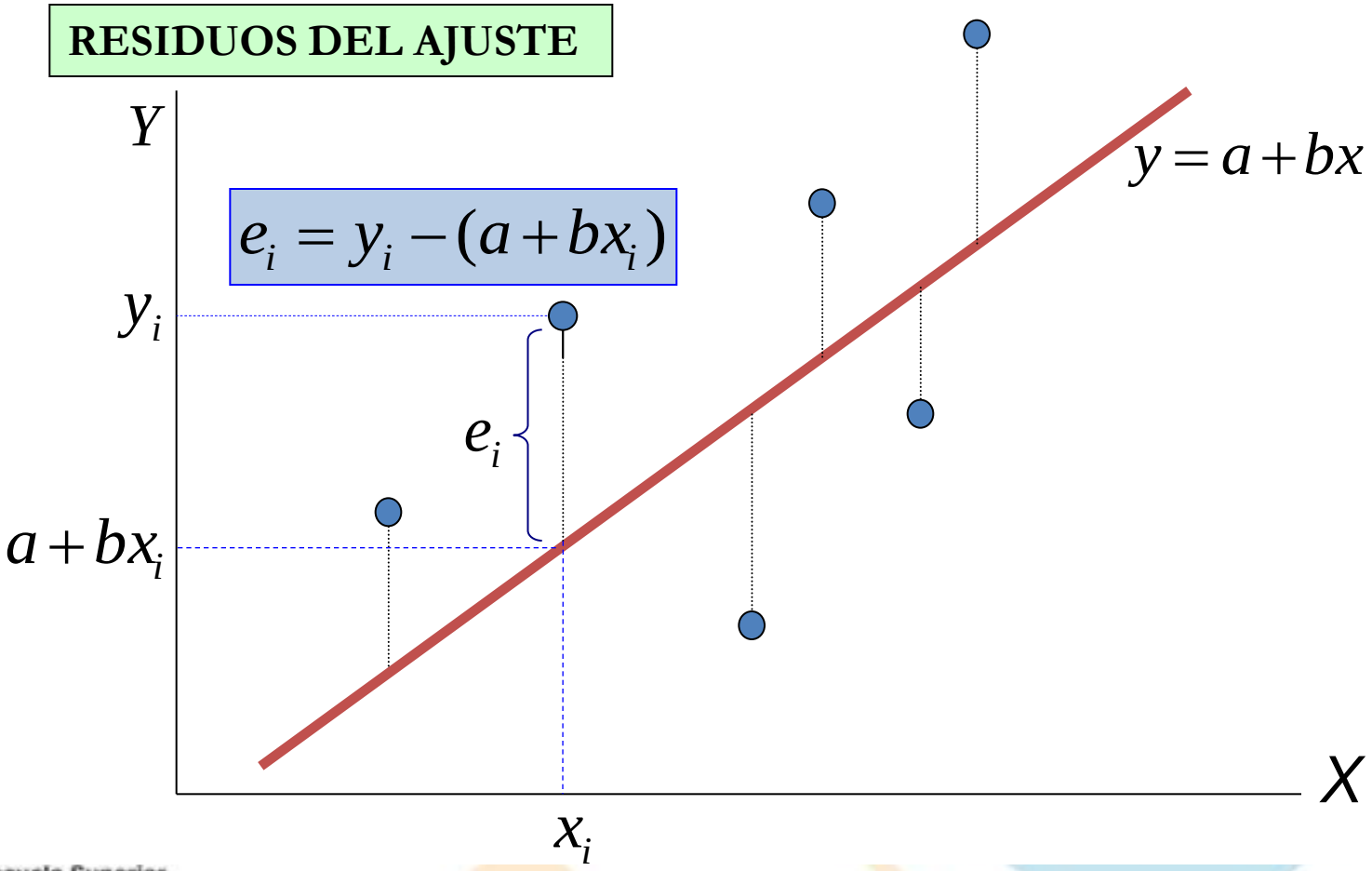
La **regresión lineal** tiene por objetivo encontrar una función que se aproxime lo mejor posible a la relación de dependencia estadística entre dos variables para predecir los valores de una de ellas en función de la otra.

Para ello es necesario determinar cuál de estas variables será la variable explicada o **variable dependiente** (por ejemplo, Y) y cuál será la variable explicativa o **variable independiente** (X).

Escuela Superior de Ingeniería

➢Cómo calcular la expresión de la recta

Supongamos una recta genérica que pase entre los puntos (x_i, y_i)



4.REGRESIÓN LINEAL

Buscamos una ecuación de una recta $y = a + bx$

que haga mínima la suma de los residuos:

$$\sum_{i=1}^{n} e_i^2 = \sum (residuos)^2$$

Es decir, buscamos *a* y *b* que hagan mínima la expresión:

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

La solución se obtiene con el llamado **método de mínimos cuadrados** para el caso lineal.

Ecuaciones normales

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i-(a+bx_i))^2=0 \quad \Rightarrow \quad \sum_{i=1}^{n}y_i=na+b\sum_{i=1}^{n}x_i$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i-(a+bx_i))^2=0 \quad \Rightarrow \quad \sum_{i=1}^{n}x_iy_i=a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2$$

Solución:

1. Parámetros (*a* y *b*) → Despejar en las ecuaciones

$$b=\frac{S_{xy}}{S_x^2} \qquad a=\bar{y}-b\bar{x}=\bar{y}-\frac{S_{xy}}{S_x^2}\bar{x}$$

2. Recta de Regresión → $$y=\bar{y}+\frac{S_{xy}}{S_x^2}(x-\bar{x})$$

4.REGRESIÓN LINEAL

➢Ejemplo. Recta de Regresión

X → Edad

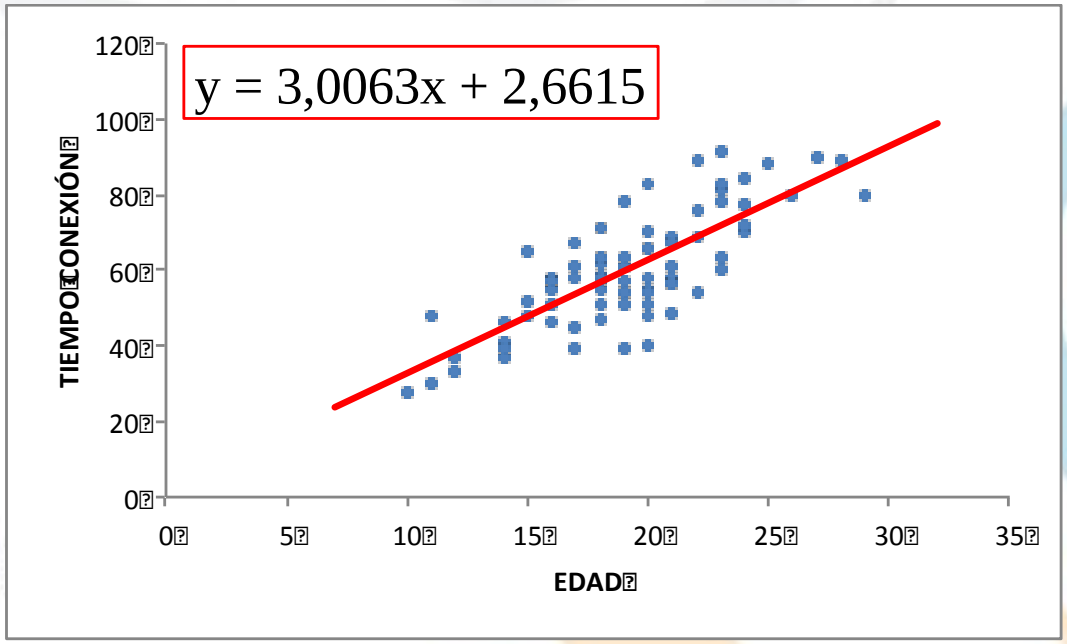
Y → Tiempo conexión

$\bar{x} = 19{,}128 \qquad S_X = 3{,}846$

$\bar{y} = 60{,}167 \qquad S_{XY} = 44{,}459$

$$b = \frac{S_{XY}}{S_X^2} = \frac{44{,}459}{3{,}846^2} = 3{,}0063$$

$$a = \bar{y} - b\bar{x} = 60{,}167 - 3{,}0063 \times 19{,}128 = 2{,}6615$$



EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

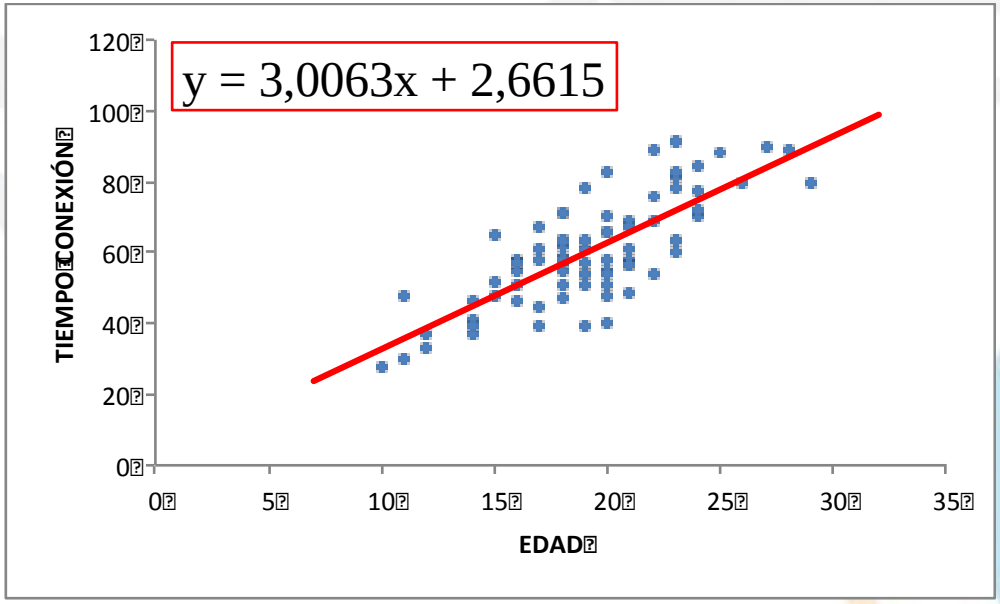
4.REGRESIÓN LINEAL

➢Ejemplo. Recta de Regresión

X → Edad
Y → Tiempo conexión

¿Cuánto tiempo de media pasará conectada una persona con 31 años?

Podemos **predecir** valores de la variable Y utilizando la ecuación de la recta de regresión



$$\hat{y} = y(31) = 3{,}0063 \times 31 + 2{,}6615 = 95{,}86 \text{ minutos}$$

EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

➢Cómo medir la bondad de la regresión

Cuanto menos dispersos sean los residuos, mejor será la bondad del ajuste

Sabiendo que: $y_i = e_i + \hat{y}_i$ y que:

Variabilidad de Y

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n};$$

Variabilidad del Error

$$S_{Error}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n};$$

Variabilidad del Modelo

$$S_{Modelo}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{n};$$

Se comprueba fácilmente que: $S_Y^2 = S_{Error}^2 + S_{Modelo}^2$

Variabilidad Total = Variabilidad del Error +Variabilidad del Modelo

El modelo lineal será mejor cuando $S_{Modelo}^2 \rightarrow S_Y^2$

Equivalentemente, $S_{error}^2 \rightarrow 0$

4.REGRESIÓN LINEAL

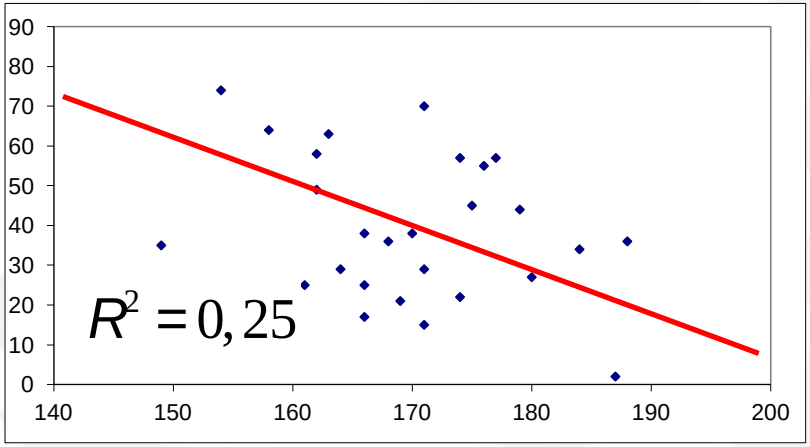
➢Cómo medir la bondad de la regresión. Coeficiente de Determinación

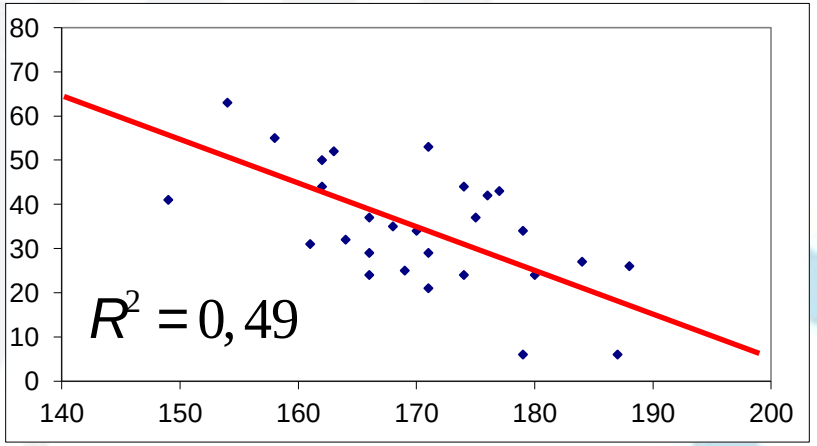
$$R^2 = \frac{\text{Variabilid ad del Modelo}}{\text{Variabilid ad Total}} = \frac{S^2_{Modelo}}{S^2_Y} = \frac{S^2_{XY}}{S^2_X S^2_Y} = \left(\frac{S_{XY}}{S_X S_Y}\right)^2 = r^2$$

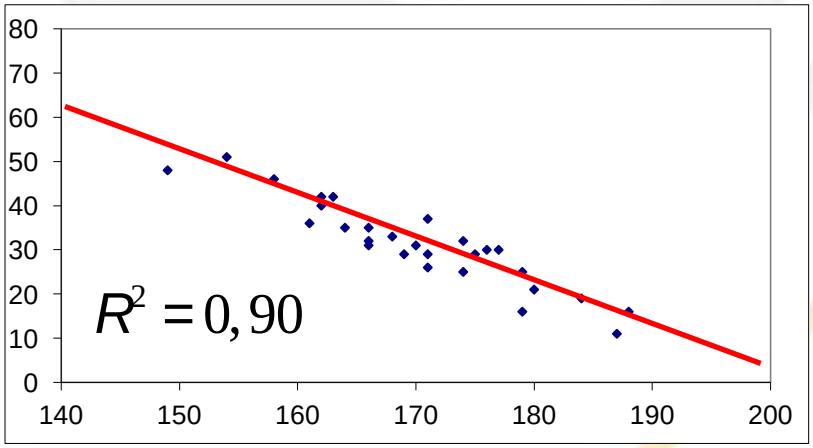
Propiedades

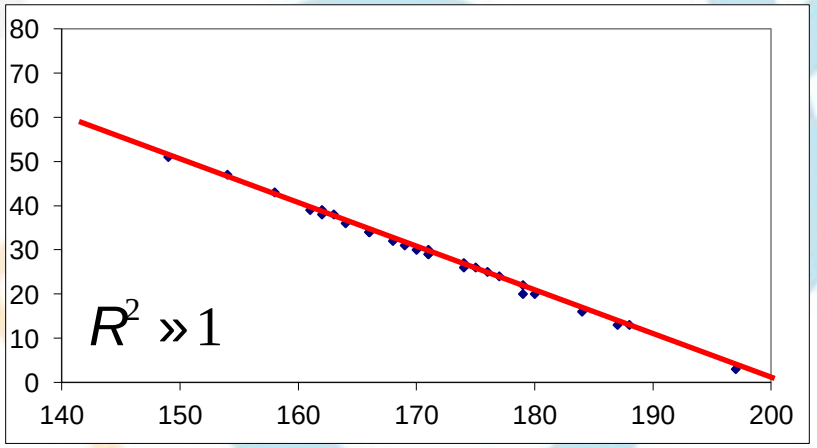
- Es adimensional
- Por construcción toma valores entre 0 y 1.
- $R^2 \cdot 100\%$ Tanto por ciento de la variabilidad total explicada por el modelo.
- Si $R^2 = 0$ → El modelo explica el 0% de la variabilidad total (el error explica el 100%) ←→ $S_{XY} = 0$ → Variables Incorreladas.
- Si $R^2 = 1$ → El modelo explica el 100% de la variabilidad total (el error explica el 0%) → La recta pasa por todos los puntos del diagrama de dispersión.
- Cuanto más próximo a 1 se encuentre R^2, mejor será el ajuste.

➢Cómo medir la bondad de la regresión. Coeficiente de Determinación









➢Ejemplo. Bondad de la Regresión

X → Edad
Y → Tiempo conexión

$\bar{x} = 19{,}128$ $\quad S_X = 3{,}846$ $\quad S_{XY} = 44{,}459$

$\bar{y} = 60{,}167$ $\quad S_Y = 14{,}950$

$$y = 2{,}6615 + 3{,}0063x \qquad r = 0'773$$

La **bondad del ajuste** de la recta de regresión viene dada por:

$$R^2 = r^2 = 0{,}773^2 = 0{,}598$$

Si multiplicamos este resultado por 100, obtenemos el porcentaje de **variabilidad de Y que queda explicada por el modelo lineal**.

En este caso, el **59,8%** de la variabilidad de Y queda explicada por el modelo lineal calculado anteriormente.

EDAD	TIEMPO CONEX.	EDAD	TIEMPO CONEX.
22	76	18	58
11	30	16	51
18	55	19	39
19	54	12	33
10	28	21	56
20	58	20	83
18	59	18	63
27	90	24	72
15	65	17	67
20	55	18	47
18	57	26	80
20	54	16	58
24	77	16	55
17	58	18	71
23	81	16	57
17	45	20	70
20	66	16	57
19	61	14	37
21	61	23	78
12	37	24	84
23	60	22	69
21	69	21	67
19	78	22	89
23	63	17	61
19	63	25	88
19	54	29	80
15	52	23	83
18	71	18	51
14	41	20	51
21	58	21	49
24	70	14	46
15	48	17	39
18	63	28	89
21	56	20	66
16	46	23	91
11	48	20	48
18	62	19	57
20	40	19	51
22	54	14	39

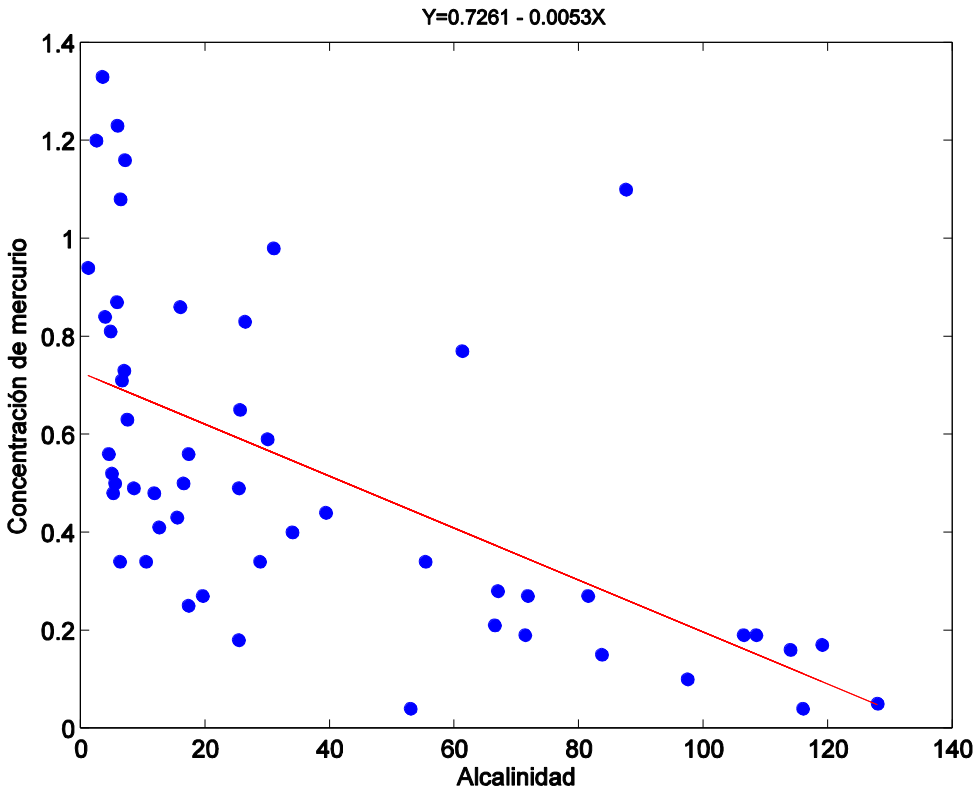
➢Observaciones sobre la Regresión

A menudo el modelo lineal f(x) = a + bx no será el que mejor describa la relación entre X e Y, o simplemente no tendrá sentido.

Ejemplo:

$$r = -0.5938;$$

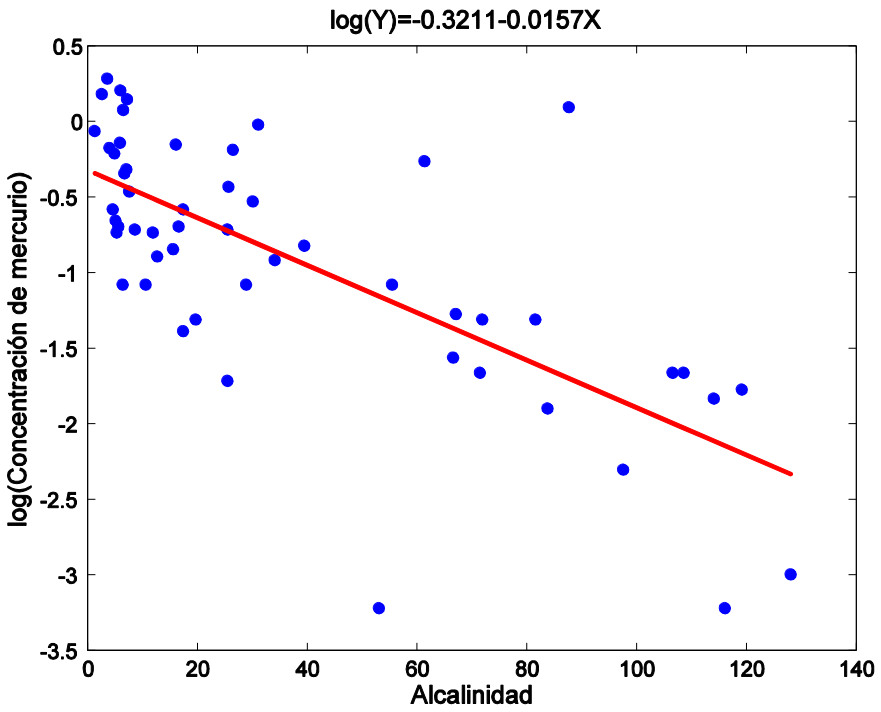
$$R^2 = 0.3527$$



➢Observaciones sobre la Regresión

Si modelizamos la relación entre X e Y incorrectamente, el modelo no dará predicciones fiables de valores desconocidos de Y en función de valores conocidos de X.

Una solución sencilla es transformar las variables Y y/o X mediante una función no lineal (logarítmo, cuadrática, exponencial ,...) y calcular la recta de regresión entre las variables transformadas.



Ejemplo:

Si tomamos el Log Y en vez de Y:

$$r_{\log} = -0.7146;$$

$$R^2 = 0.51$$

EJERCICIO 2

En el estudio de la deshidratación de un derivado industrial, la variable X indica la cantidad de agua y la variable Y la presión en atmósferas a la que se somete el compuesto. La siguiente tabla ofrece los datos tomados tras realizar 10 pruebas:

X	253	232	210	200	191	187	134	102	81	25
Y	2	3	4	5	6	7	8	9	10	11

(a) ¿Entre qué valores se encuentra el 50% de las presiones centrales?
(b) Asumiendo que existe una relación lineal entre ambas variables, ¿qué cantidad de agua cabe esperar para una presión de 6.5 atmósferas? ¿Es fiable esa predicción?

Sol. (a) Q1 = 4 y Q3 = 9; (b) 161;5; Es fiable.