Erregresio linealaren analisia

Datuak prozesatzeko sistema baten oinarrizko egitura-elementuak hiru dira:

- **Fitxategiak**. Sisteman dauden erregistro-multzo iraunkor bat osatzen dute.
- **Fluxuak**. Sistema eta inguruaren arteko datu-interfazeak dira.
- **Prozesuak**. Funtzionalki definitutako datuen manipulazio logikoak.

Hiru elementu hauei buruzko ikerketa bati egiten zaio aurre software-n garapenaren **kostuak** ikertzeko asmoz.

Datuak prozesatzeko sistema baten oinarrizko egitura-elementuak hiru dira:

- **Fitxategiak**. Sisteman dauden erregistro-multzo iraunkor bat osatzen dute.
- **Fluxuak**. Sistema eta inguruaren arteko datu-interfazeak dira.
- **Prozesuak**. Funtzionalki definitutako datuen manipulazio logikoak.

Hiru elementu hauei buruzko ikerketa bati egiten zaio aurre software-n garapenaren **kostuak** ikertzeko asmoz.

Datuak hauek izan ziren:

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

Datuak prozesatzeko sistema baten oinarrizko egitura-elementuak hiru dira:

- **Fitxategiak.** Sisteman dauden erregistro-multzo iraunkor bat osatzen dute.
- **Fluxuak.** Sistema eta inguruaren arteko datu-interfazeak dira.
- **Prozesuak.** Funtzionalki definitutako datuen manipulazio logikoak.

Hiru elementu hauei buruzko ikerketa bati egiten zaio aurre software-n garapenaren **kostuak** ikertzeko asmoz.

Datuak hauek izan ziren:

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

Fitxategiak (X_1), **Fluxuak** (X_2), eta **Prozesuak** (X_3), aldagai *askeak* dira.

Kostuak (Y), *menpeko* aldagaia.

Datu-taula

Ω, objektuak — X_j, aldagai ask*eak*

Y, *menpeko* aldagaia

$X_j(\omega_i)=x_{ij} \quad i=1,2,\ldots,n \qquad j=1,2,\ldots,p$

$Y(\omega_i)=y_i \quad i=1,2,\ldots,n$

Datu-taula

Ω, objektuak — X_j, aldagai ask*eak*
Y, *menpeko* aldagaia

$X_j(\omega_i)=x_{ij} \quad i=1,2,\ldots,n \qquad j=1,2,\ldots,p$

$Y(\omega_i)=y_i \quad i=1,2,\ldots,n$

Ω	X_1		X_j		X_p	Y
ω_1	x_{11}		x_{1j}		x_{1p}	y_1
ω_i	x_{i1}		x_{ij}		x_{ip}	y_i
ω_n	x_{n1}		x_{nj}		x_{np}	y_n

Datu-taula

X_j ,   aldagai askeak, *kuantitatiboak*
Y,   menpeko aldagaia, *kuantitatiboa*

Datu-taula

X_j , aldagai askeak, *kuantitatiboak*
Y, menpeko aldagaia, *kuantitatiboa*

$X_j(\omega_i)=x_{ij}$ *zenbakiak* $Y(\omega_i)=y_i$ *zenbakiak*

Datu-taula

X_j, aldagai askeak, *kuantitatiboak*
Y, menpeko aldagaia, *kuantitatiboa*

$X_j(\omega_i)=x_{ij}$ *zenbakiak* $Y(\omega_i)=y_i$ *zenbakiak*

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

***X*ekiko *Y*ren erregresioa**

X_j , aldagai askeak, *kuantitatiboak*
Y, menpeko aldagaia, *kuantitatiboa*

$X_j(\omega_i)=x_{\mathrm{ij}}$ *zenbakiak* $Y(\omega_i)=y_i$ *zenbakiak*

Erregresioa (***X***ekiko ***Y***ren erregresioa):

$$Y = f(X_1, X_2, \ldots, X_j, \ldots, X_p) + E$$

← E, errorea, akatsa, hondarra

$$Y(\omega_i) = f(X_1(\omega_i), X_2(\omega_i), \ldots, X_j(\omega_i), \ldots, X_p(\omega_i)) + E(\omega_i)$$

Erregresio lineala

$$\boldsymbol{Y = a_0+a_1X_1+a_2X_2+...a_jX_j+...+a_pX_p + E}$$

$$Y(\omega_i) = a_0+a_1X_1(\omega_i)+a_2X_2(\omega_i)+...a_jX_j(\omega_i)+...+a_pX_p(\omega_i) + E(\omega_i)$$

Erregresio lineala:

$$\boldsymbol{Y = a_0+a_1X_1+a_2X_2+...a_jX_j+...+a_pX_p + E}$$

$$Y(\omega_i) = a_0+a_1X_1(\omega_i)+a_2X_2(\omega_i)+...a_jX_j(\omega_i)+...+a_pX_p(\omega_i) + E(\omega_i)$$

Optimizazio-problema:

Kalkulatu a_j *(j=0,1,2,...,p) hoberenak*, **errorea** (oro har) ahalik eta txikiena izan dadin.

Erregresio lineala:

$$\boldsymbol{Y = a_0+a_1X_1+a_2X_2+...a_jX_j+...+a_pX_p + E}$$

$$Y(\omega_i) = a_0+a_1X_1(\omega_i)+a_2X_2(\omega_i)+...a_jX_j(\omega_i)+...+a_pX_p(\omega_i) + E(\omega_i)$$

Optimizazio-problema:

Kalkulatu a_j *(j=0,1,2,...,p) hoberenak*, **errorea** (oro har) ahalik eta txikiena izan dadin.

Hoberenak, **errore karratu txikien**en zentzuan:

$$\frac{1}{n}\sum_{i=1}^{n} E^2(\omega_i)$$

Erregresio lineala:

$$\boldsymbol{Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots a_j X_j + \ldots + a_p X_p + E}$$

$$Y(\omega_i) = a_0 + a_1 X_1(\omega_i) + a_2 X_2(\omega_i) + \ldots a_j X_j(\omega_i) + \ldots + a_p X_p(\omega_i) + E(\omega_i)$$

Optimizazio-problema:

Kalkulatu a_j *(j=0,1,2,...,p) hoberenak*, **errorea** (oro har) ahalik eta txikiena izan dadin.

Hoberenak, **errore karratu txikien**en zentzuan:

$$\frac{1}{n}\sum_{i=1}^{n} E^2(\omega_i)$$

$\frac{1}{n}$ -k ez du eraginik, baina batezbestekoaren ulerkera ematen dio

Erregresio lineal bakuna

(aldagai *aske* bakarra)

$$\boldsymbol{Y = b + aX + E}$$

$$Y(\omega_i) = b + aX(\omega_i) + E(\omega_i)$$

$$Y'(\omega_i) = b + aX(\omega_i)$$

$$E(\omega_i) = Y(\omega_i) - Y'(\omega_i)$$

Erregresio lineal bakuna

(aldagai *aske* bakarra)

$$\boldsymbol{Y = b + aX + E}$$

$$Y(\omega_i) = b + aX(\omega_i) + E(\omega_i)$$

$$Y'(\omega_i) = b + aX(\omega_i) \qquad\qquad E(\omega_i) = Y(\omega_i) - Y'(\omega_i)$$

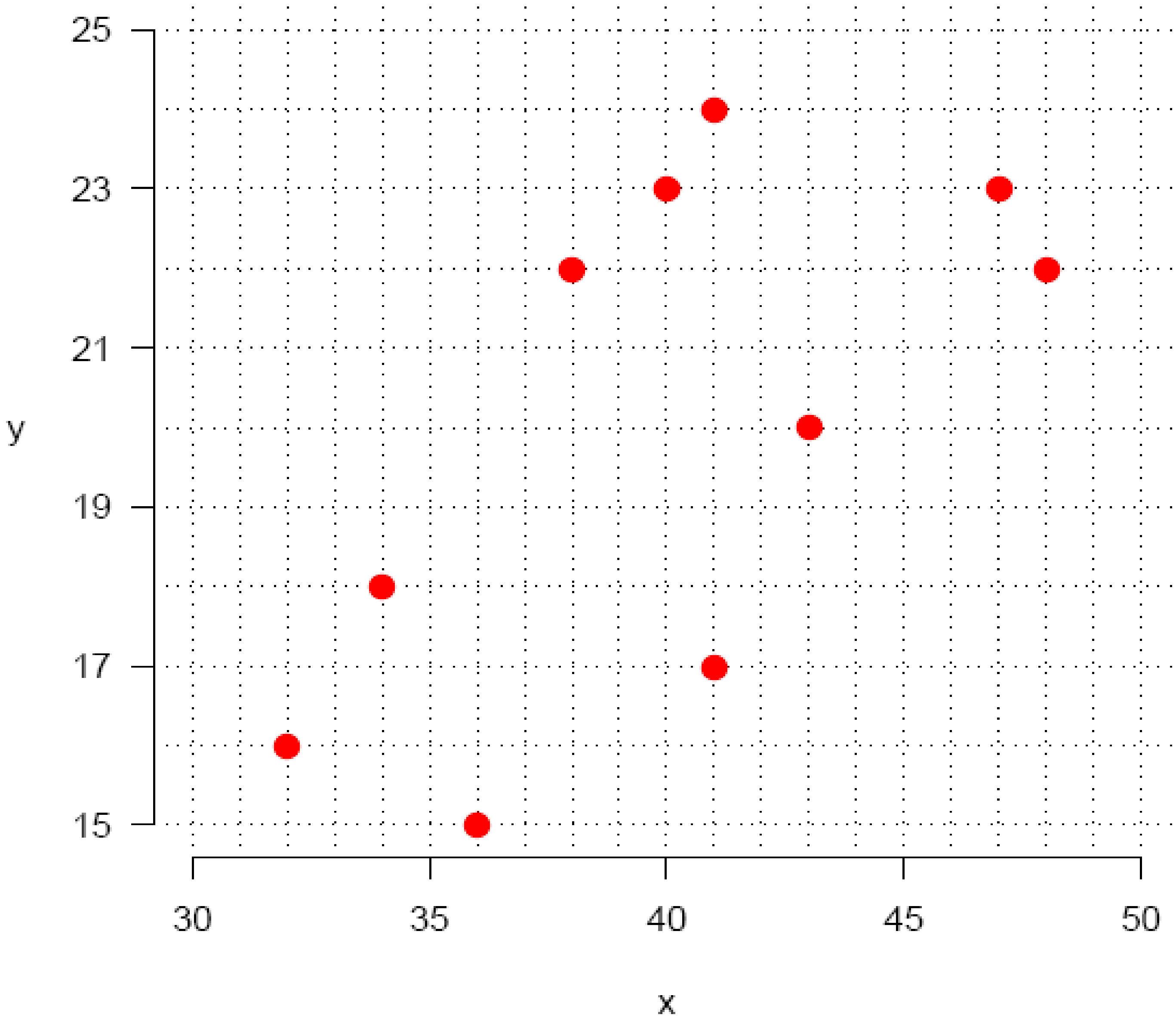
Optimizazio-problema:

Kalkulatu *b* eta *a hoberenak:*

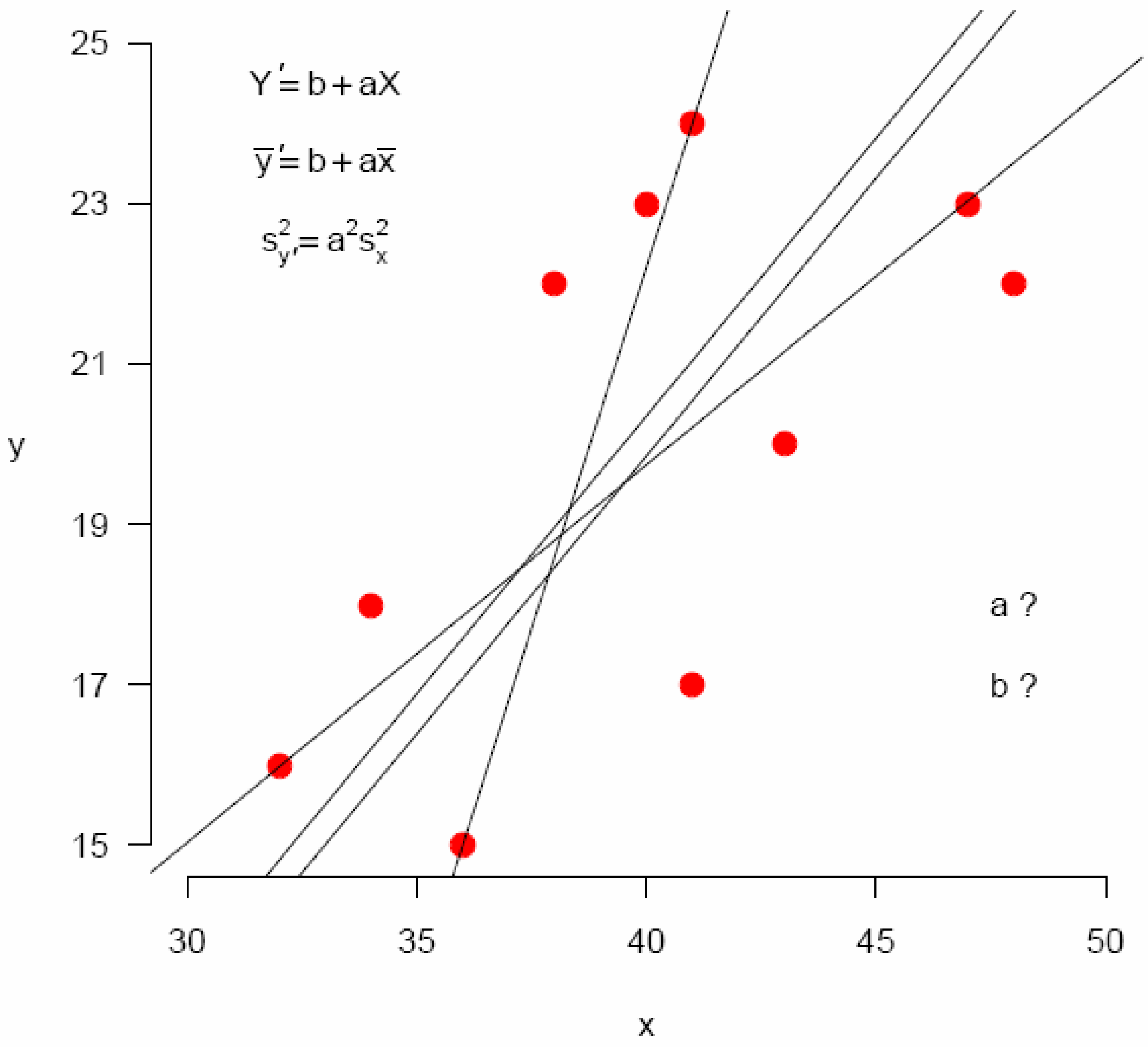
$$\underset{b,\,a}{\text{Minimizatu}} \quad \frac{1}{n}\sum_{i=1}^{n} E^2(\omega_i) = \frac{1}{n}\sum_{i=1}^{n}(Y(\omega_i) - Y'(\omega_i))^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y(\omega_i) - b - aX(\omega_i))^2$$

X	41	38	48	32	34	36	41	43	47	40
Y	17	22	22	16	18	15	24	20	23	23



Zein zuzen da egokiena ***Y*ren erregresioa** adierazteko**?**



*X*rekiko *Y*ren erregresio lineal bakuna

$\boldsymbol{Y = b + aX + E}$

Notazioa arinduz: $X(\omega_i)$, x_i, $Y(\omega_i)$, y_i , $Y'(\omega_i)$, y'_i, $E(\omega_i)$, e_i

$y_i = b + a \cdot x_i + e_i$ $y'_i = b + a \cdot x_i$ $e_i = y_i - y'_i$

Xrekiko Yren erregresio lineal bakuna

$\boldsymbol{Y = b + aX + E}$

Notazioa arinduz: $X(\omega_i), x_i, \ Y(\omega_i), y_i, \ Y'(\omega_i), y'_i, \quad E(\omega_i), e_i$

$y_i = b + a \cdot x_i + e_i \qquad y'_i = b + a \cdot x_i \qquad e_i = y_i - y'_i$

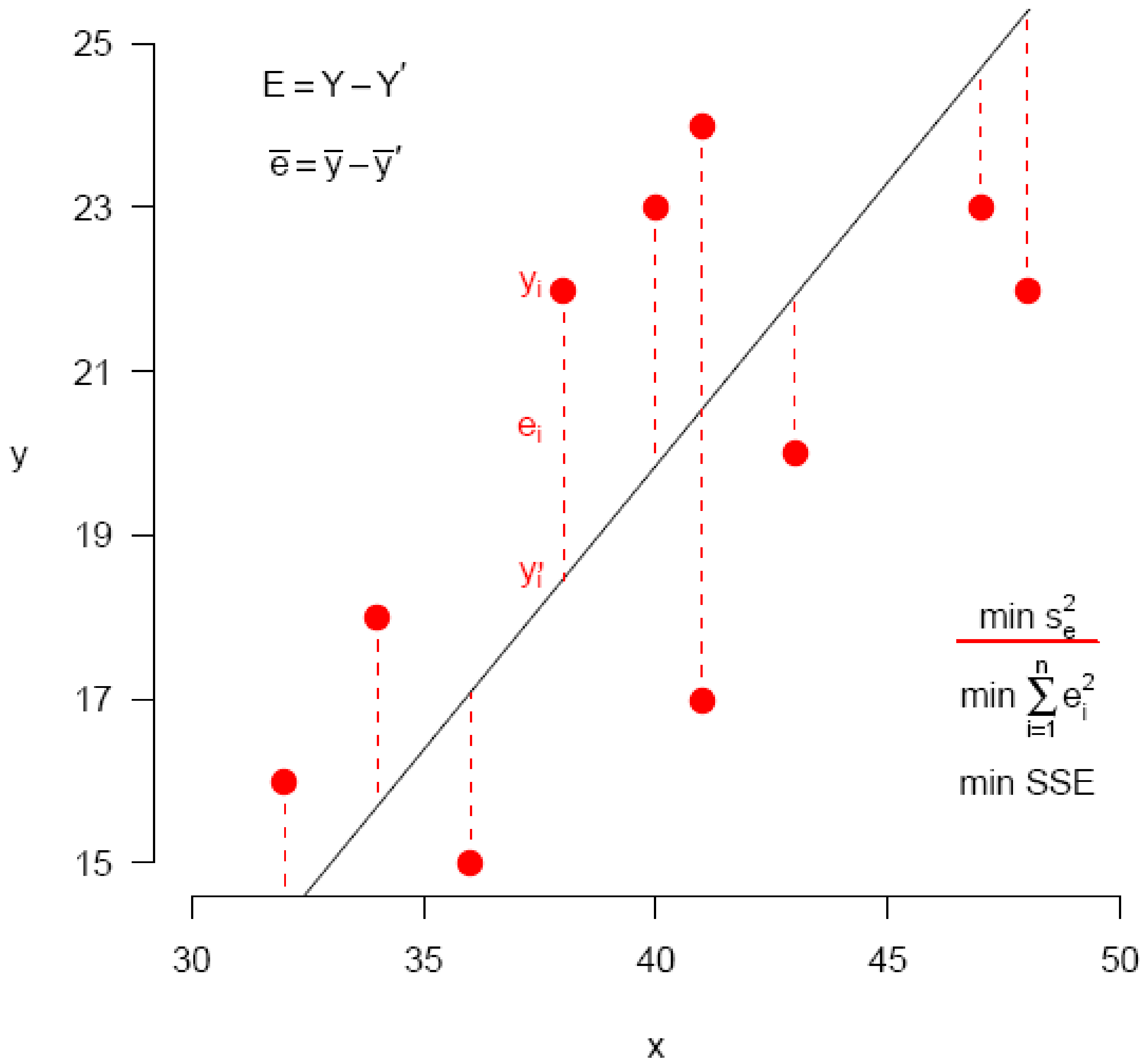
Optimizazio-problema:

Kalkulatu b eta a *hoberenak:*

$$\underset{b,\,a}{\text{Minimizatu}} \quad \frac{1}{n}\sum_{i=1}^{n} e^2_i = \frac{1}{n}\sum_{i=1}^{n} (y_i - y'_i)^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - b - a \cdot x_i)^2$$

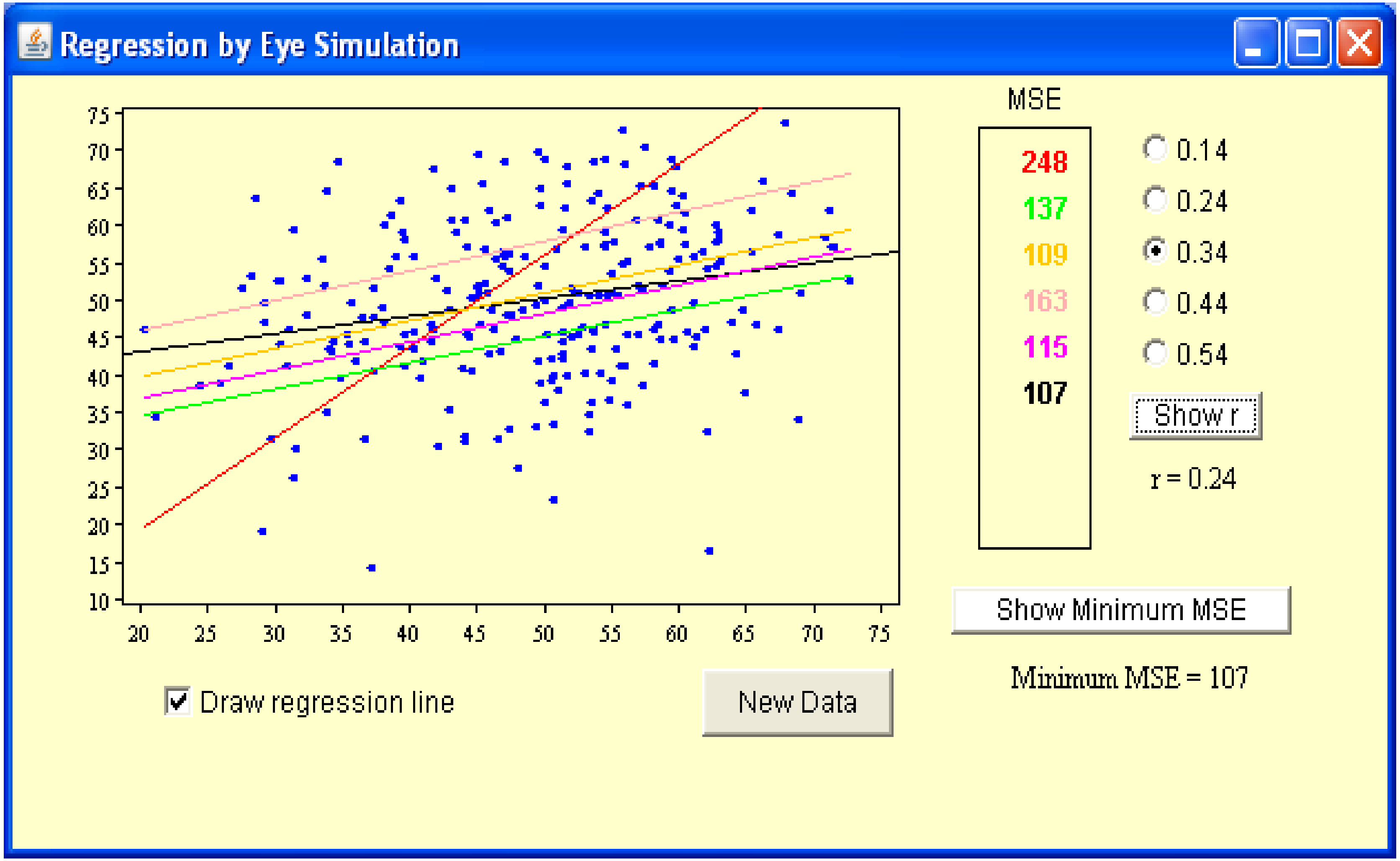
$$\underset{b,\,a}{\text{Minimizatu}} \quad G(a,b) = \sum_{i=1}^{n} (y_i - b - a \cdot x_i)^2$$

*X*rekiko *Y*ren erregresio lineal bakuna



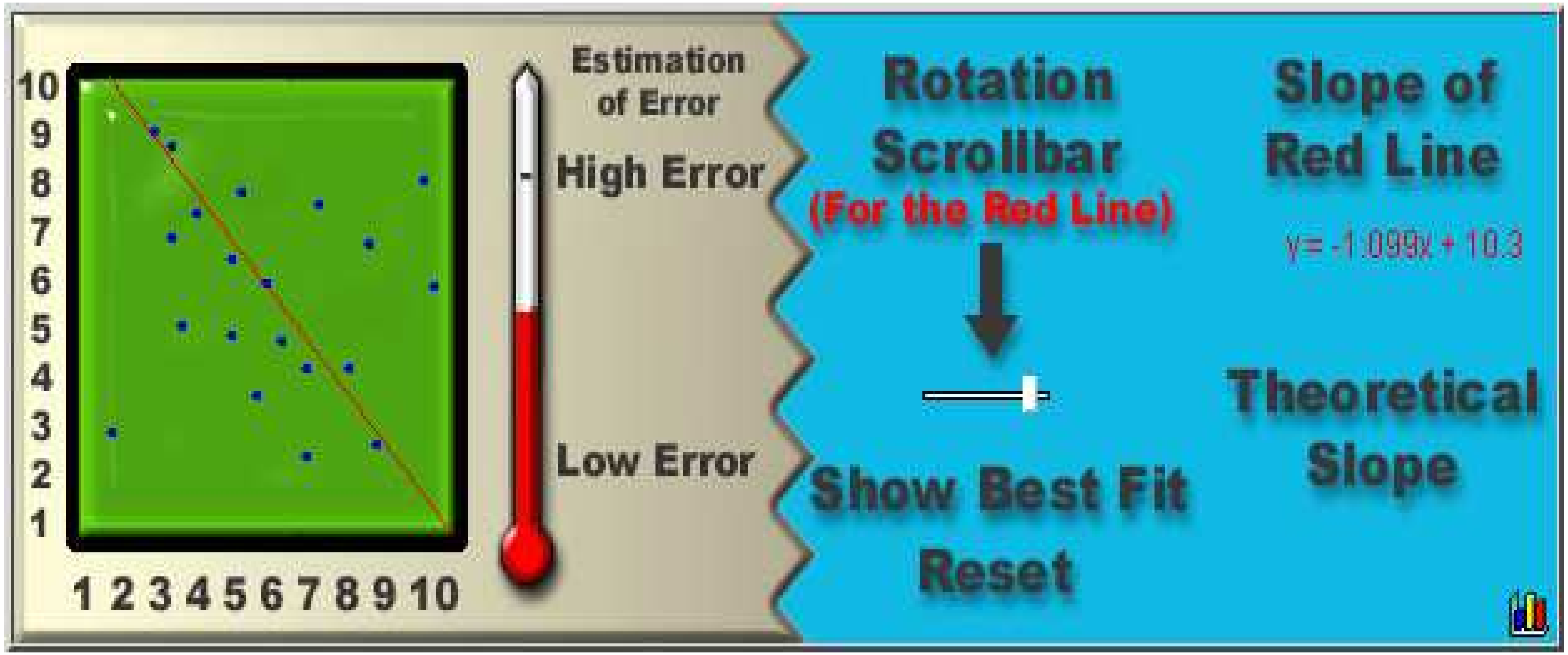
*X*rekiko *Y*ren erregresio lineal bakuna

http://onlinestatbook.com/stat_sim/reg_by_eye/index.html



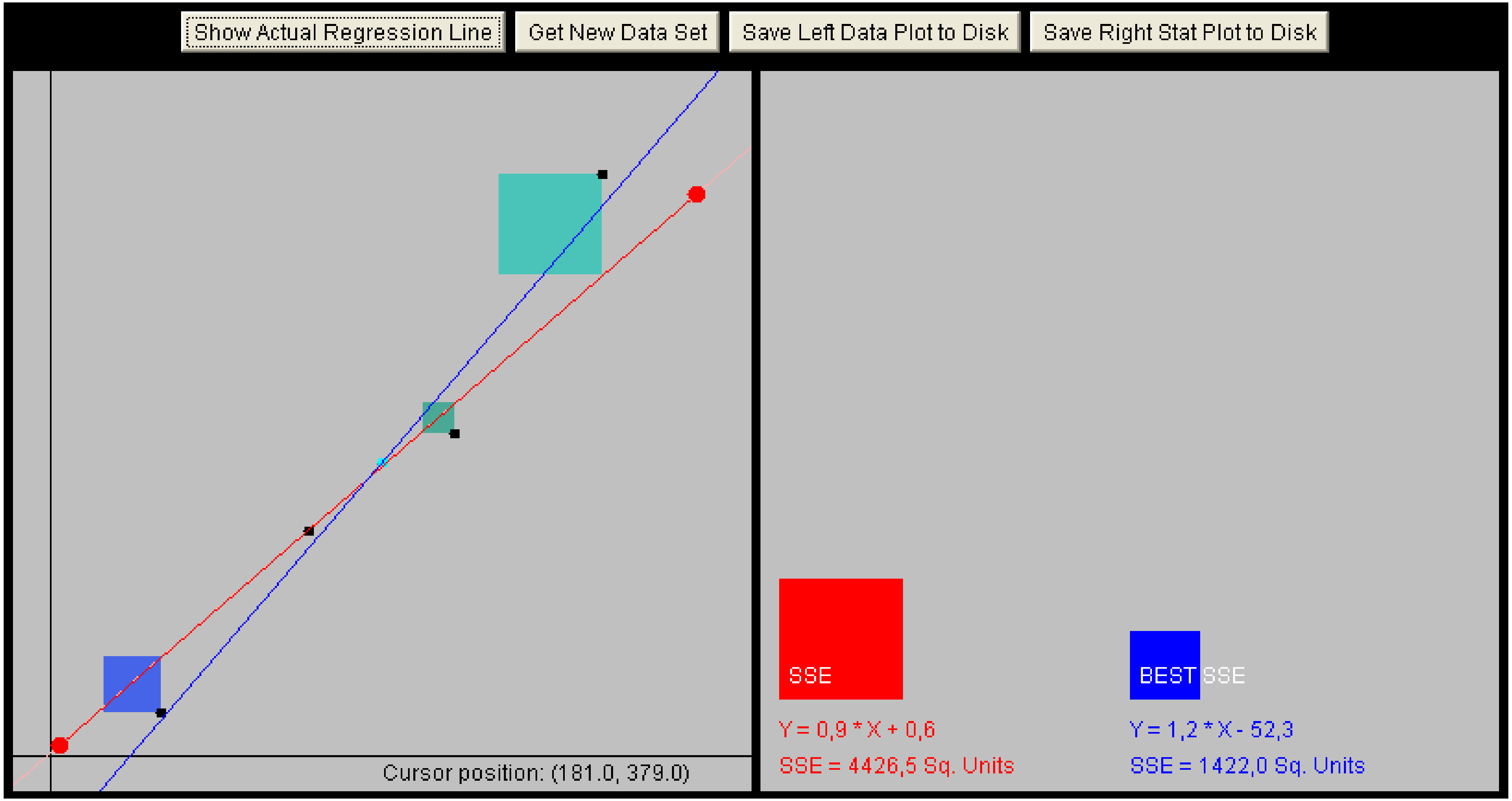
*X*rekiko *Y*ren erregresio lineal bakuna

http://www.mste.uiuc.edu/activity/regression/



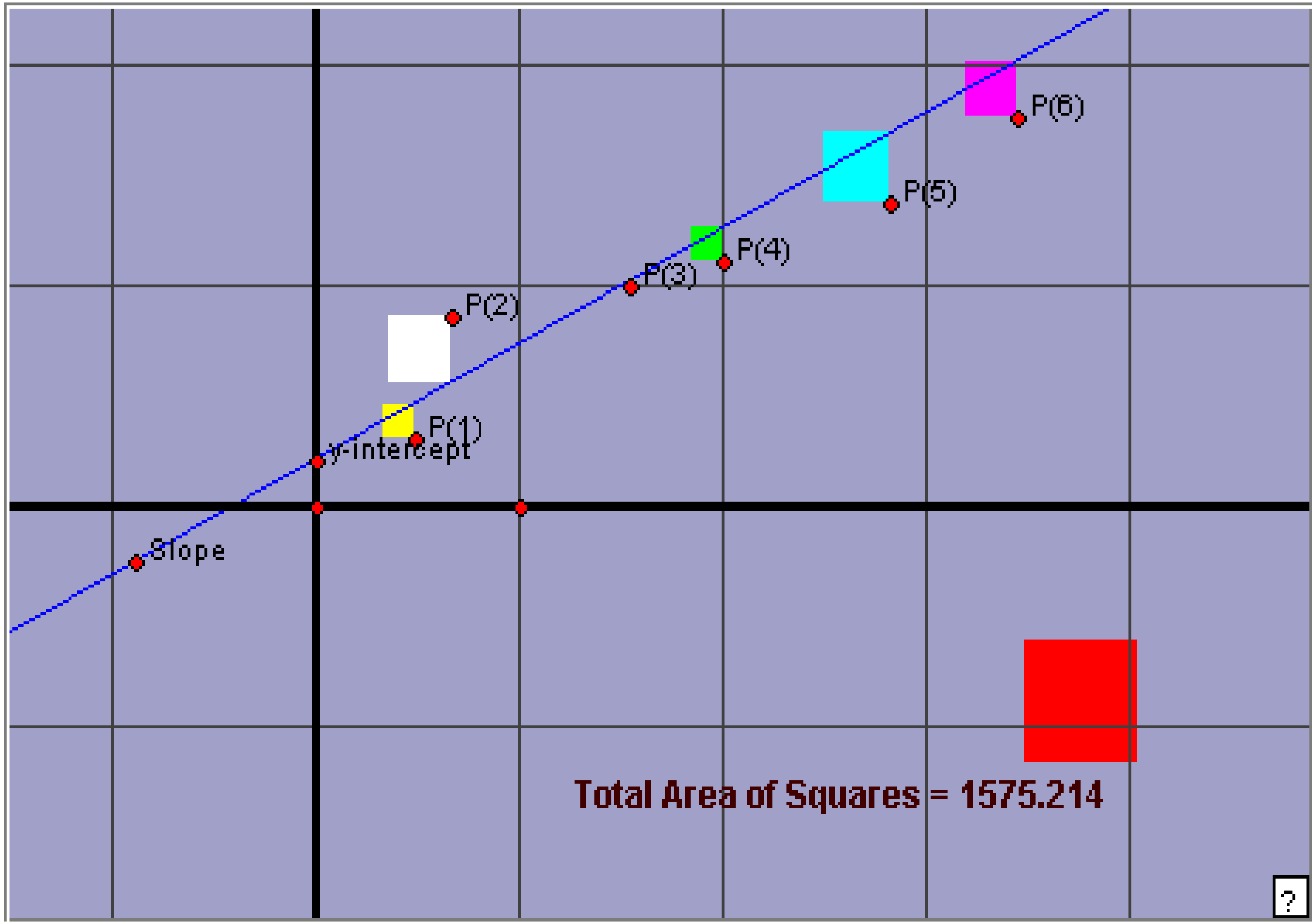
*X*rekiko *Y*ren erregresio lineal bakuna

http://www.math.wpi.edu/Course_Materials/SAS/lablets/7.3/7.3a/index.html



*X*rekiko *Y*ren erregresio lineal bakuna

http://www.dynamicgeometry.com/JavaSketchpad/Gallery/Other_Explorations_and_Amusements/Least_Squares.html



*X*rekiko *Y*ren erregresio lineal bakuna

$$\boldsymbol{Y = b + aX + E}$$

Optimizazio-problema:

$$\underset{b,\,a}{\text{Minimizatu}} \quad G(a,b) = \sum_{i=1}^{n} (y_i - b - a \cdot x_i)^2$$

*X*rekiko *Y*ren erregresio lineal bakuna

$$Y = b + aX + E$$

Optimizazio-problema:

$$\underset{b,\, a}{\text{Minimizatu}} \quad G(a,b)=\sum_{i=1}^{n}(y_i-b-a\cdot x_i)^2$$

<u>Ekuazio linealak</u>:

$$\delta G(a,b)/\delta a=\sum_{i=1}^{n}2\cdot(y_i-\mathbf{b}-\mathbf{a}\cdot x_i)(-x_i) \quad = 0$$

$$\delta G(a,b)/\delta b=\sum_{i=1}^{n}2\cdot(y_i-\mathbf{b}-\mathbf{a}\cdot x_i)(-1) \quad = 0$$

*X*rekiko *Y*ren erregresio lineal bakuna

$$\boldsymbol{Y = b + aX + E}$$

Optimizazio-problema:

$$\underset{b,\,a}{Minimizatu} \quad G(a,b) = \Sigma_{i=1}^{n} (y_i - b - a \cdot x_i)^2$$

Ekuazio linealak:

$$\delta G(a,b)/\delta a = \Sigma_{i=1}^{n} 2 \cdot (y_i - \boldsymbol{b} - \boldsymbol{a} \cdot x_i)(-x_i) \quad = 0$$

$$\delta G(a,b)/\delta b = \Sigma_{i=1}^{n} 2 \cdot (y_i - \boldsymbol{b} - \boldsymbol{a} \cdot x_i)(-1) \quad = 0$$

Soluzioa:

$$\boldsymbol{b} = \overline{y} - \boldsymbol{a} \cdot \overline{x} \qquad \boldsymbol{a} = \frac{\Sigma_{i=1}^{n} x_i \cdot y_i - n \cdot \overline{x} \cdot \overline{y}}{\Sigma_{i=1}^{n} x_i^2 - n \cdot \overline{x}^2}$$

*X*rekiko *Y*ren erregresio lineal bakuna

Soluzioaren ulerkera

X aldagaiaren *bariantza*:

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

*X*rekiko *Y*ren erregresio lineal bakuna

Soluzioaren ulerkera

X aldagaiaren *bariantza*:

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

Era berean: $\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x}\cdot\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})$

*X*rekiko *Y*ren erregresio lineal bakuna

Soluzioaren ulerkera

X aldagaiaren *bariantza*:

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

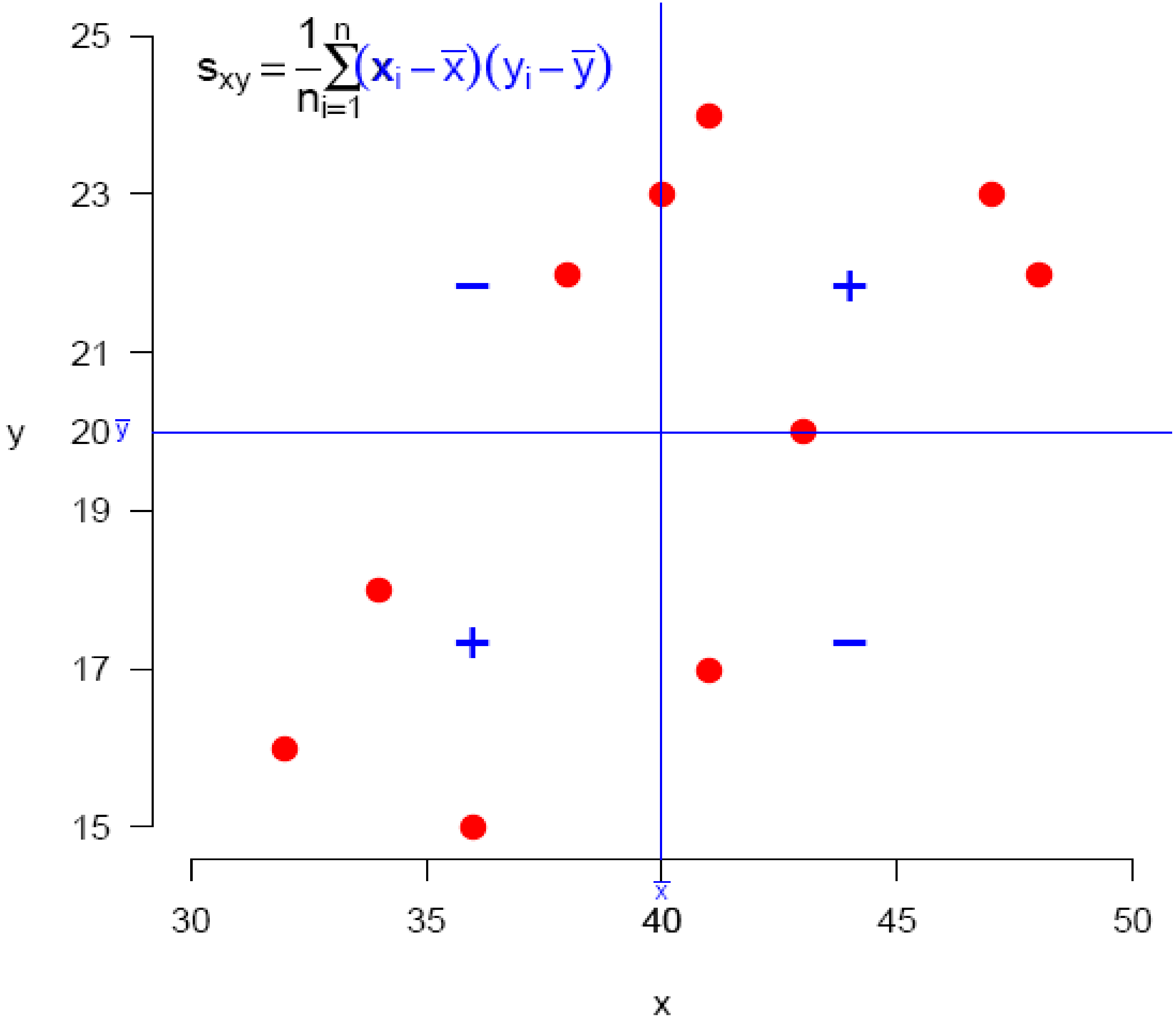
Era berean: $\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x} \cdot \overline{y} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})$

Honi ***kobariantza*** esaten zaio:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \qquad s_{xx} = s_x^2$$

***X*rekiko *Y*ren erregresio lineal bakuna**

Kobariantzaren ulerkera bat



***X*rekiko *Y*ren erregresio lineal bakuna**

Soluzioaren ulerkera

X aldagaiaren *bariantza*:

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

Era berean: $\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i - \overline{x}\cdot\overline{y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})$

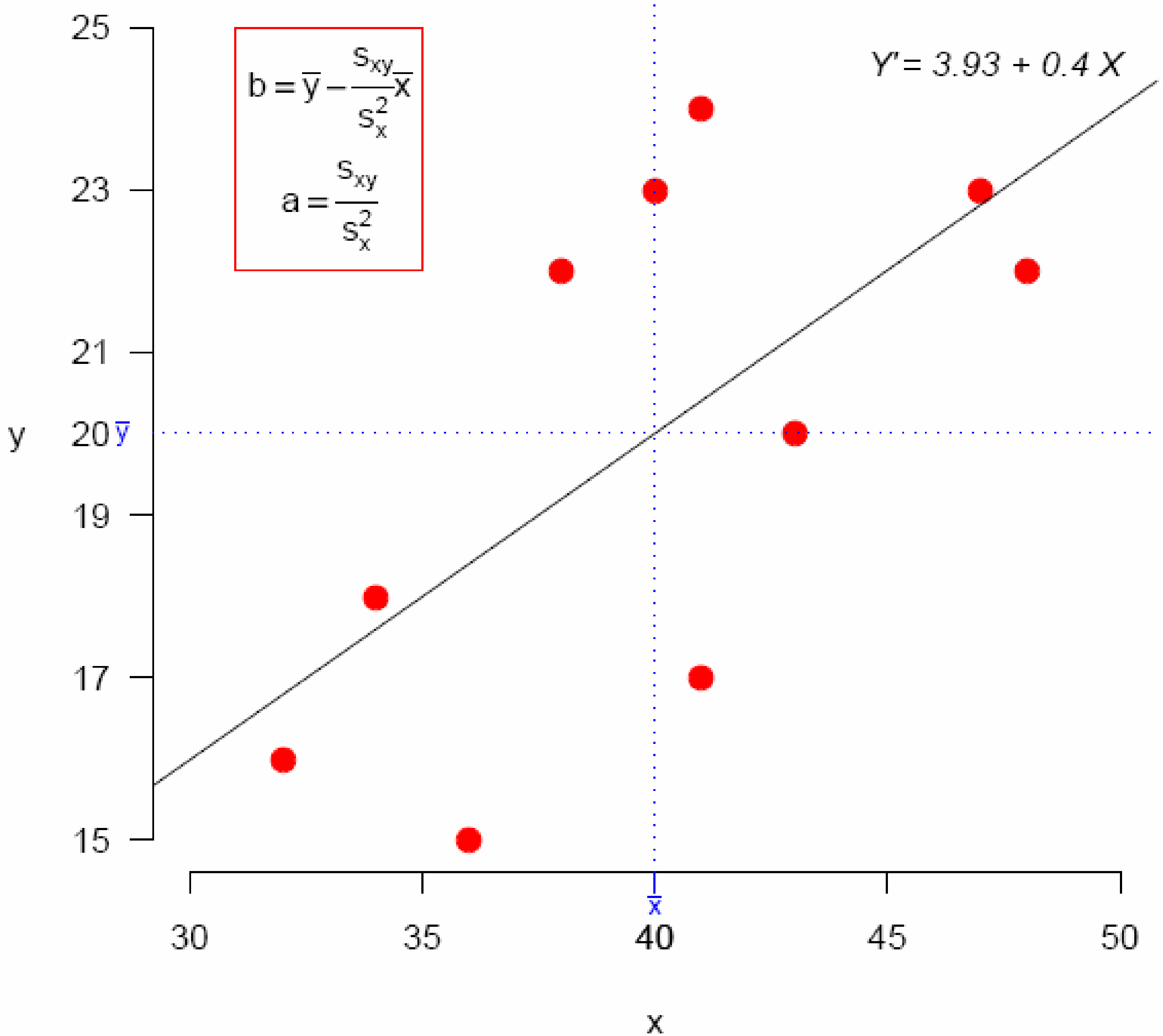
Honi ***kobariantza*** esaten zaio:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y}) \qquad s_{xx} = s_x^2$$

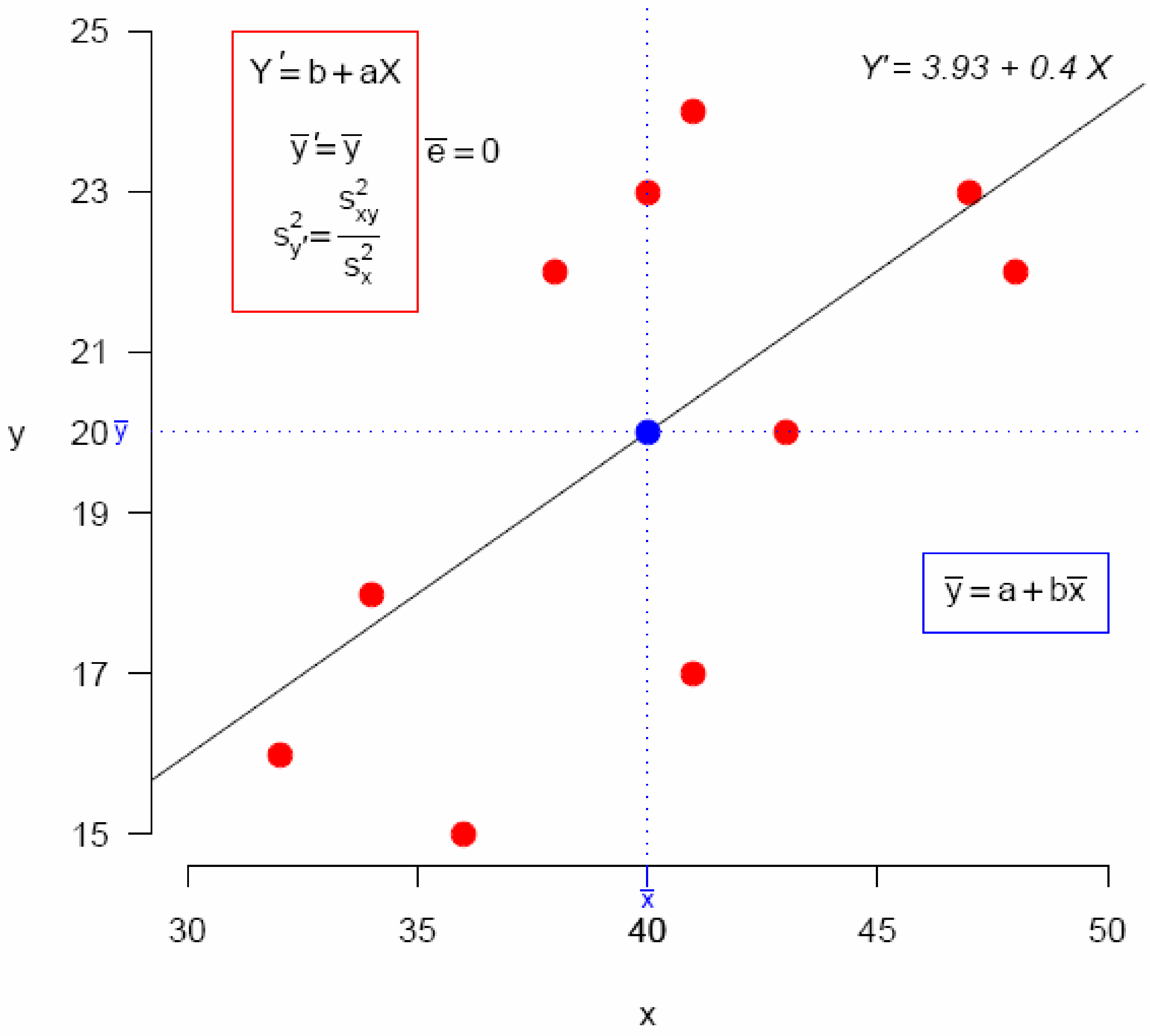
Horrela, *optimizazio-problemaren <u>soluzioa</u>*:

$$b = \overline{y} - a\cdot\overline{x} \qquad a = \frac{s_{xy}}{s_x^2}$$

*X*rekiko *Y*ren erregresio lineal bakuna



*X*rekiko *Y*ren erregresio lineal bakuna



*X*rekiko *Y*ren erregresio lineal bakuna

Y' = b+aX

$$y' = b + a\cdot\overline{x} = \left(\overline{y} - \frac{s_{xy}}{s_x^2}\cdot\overline{x}\right) + \frac{s_{xy}}{s_x^2}\cdot\overline{x} = \overline{y}$$

$(\overline{x}, \overline{y})$ Erregresio-zuzenaren puntu bat da

$$Y' = 3.93 + 0.40\cdot X$$

$$y' = 3.9344 + 0.4016\cdot\overline{x} = 3.9344 + 0.4016\cdot 40 = 20 = \overline{y}$$

*X*rekiko *Y*ren erregresio lineal bakuna

$$\boldsymbol{Y' = b + aX} \qquad a = \frac{s_{xy}}{s_x^2}$$

$$s_{y'}^2 = a^2 \cdot s_x^2 = \frac{s_{xy}^2}{s_x^2}$$

$$s_{xy'} = a \cdot s_x^2 = s_{xy}$$

$$s_{yy'} = a \cdot s_{xy} = \frac{s_{xy}^2}{s_x^2} = s_{y'}^2$$

*X*rekiko *Y*ren erregresio lineal bakuna

$$Y' = b + aX \qquad a = \frac{s_{xy}}{s_x^2}$$

$$s_{y'}^2 = a^2 \cdot s_x^2 = \frac{s_{xy}^2}{s_x^2}$$

$$s_{xy'} = a \cdot s_x^2 = s_{xy}$$

$$s_{yy'} = a \cdot s_{xy} = \frac{s_{xy}^2}{s_x^2} = s_{y'}^2$$

$$E = Y - Y' \qquad e_i = y_i - y'_i$$

$$\boxed{\bar{e} = \bar{y} - \bar{y}' = 0}$$

$$s_e^2 = \Sigma_{i=1}^{n} (y_i - y'_i)^2 =$$

$$= s_y^2 + s_{y'}^2 - 2 \cdot s_{yy'}$$

$$\boxed{s_e^2 = s_y^2 - s_{y'}^2}$$

*X*rekiko *Y*ren erregresio lineal bakuna

Optimizazio-problema balio tipikoen bitartez azalduz gero, soluzio bera lortzen da, *noski*.

*X*rekiko *Y*ren erregresio lineal bakuna

Optimizazio-problema balio tipikoen bitartez azalduz gero, soluzio bera lortzen da, noski.

Balio tipikoak:

$$z_x = \frac{X - \bar{x}}{s_x} \qquad \bar{z}_x = 0 \quad s_{zx} = 1$$

$$z_y = \frac{Y - \bar{y}}{s_y} \qquad \bar{z}_y = 0 \quad s_{zy} = 1$$

$$s_{zxzy} = \frac{s_{xy}}{s_x s_y} \longleftarrow r$$

$$a^* = \frac{s_{zxzy}}{s_{zx}^2} = r \qquad b^* = \bar{z}_y - a \cdot \bar{z}_x = 0$$

$$\frac{Y - \bar{y}}{s_y} = r \cdot \frac{X - \bar{x}}{s_x} + E^*$$

$$Y = r \cdot \frac{s_y}{s_x} X + (\bar{y} - r \cdot \frac{s_y}{s_x} \cdot \bar{x}) + E^* \frac{s_y}{s_x} \qquad E = E^* \frac{s_y}{s_x}$$

Xrekiko Yren erregresio lineal bakuna

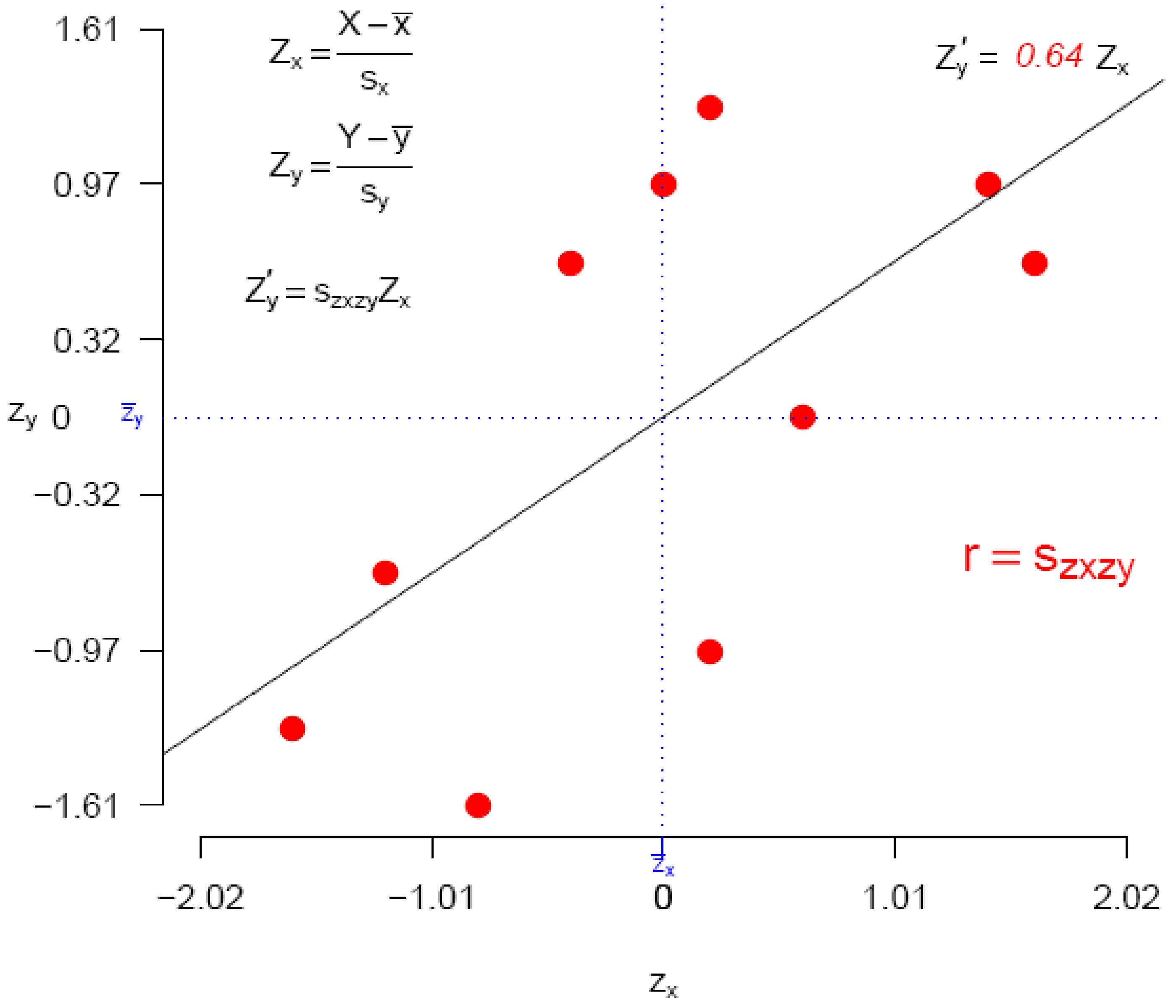
Optimizazio-problema balio tipikoen bitartez azalduz gero, soluzio bera lortzen da, noski.

$$Y = r\cdot\frac{s_y}{s_x}X+(\overline{y}-r\cdot\frac{s_y}{s_x}\cdot\overline{x})+E$$

$$Y = \frac{s_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x}X+(\overline{y}-\frac{s_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x}\cdot\overline{x})+E$$

$$Y = \frac{s_{xy}}{s_x^2}\cdot X+(\overline{y}-\frac{s_{xy}}{s_x^2}\cdot\overline{x})+E$$

*X*rekiko *Y*ren erregresio lineal bakuna



Ω	X	Y	X^2	Y^2	$X \cdot Y$	Y'	E_Y	Y'^2	E_Y^2
ω_1	41	17	1681	289	697	20.40	-3.40	416.28	11.58
ω_2	38	22	1444	484	836	19.20	2.80	368.56	7.85
ω_3	48	22	2304	484	1056	23.22	-1.22	538.94	1.48
ω_4	32	16	1024	256	512	16.79	-0.79	281.84	0.62
ω_5	34	18	1156	324	612	17.59	0.41	309.48	0.17
ω_6	36	15	1296	225	540	18.40	-3.39	338.38	11.53
ω_7	41	24	1681	576	984	20.40	3.60	416.28	12.94
ω_8	43	20	1849	400	860	21.21	-1.21	449.69	1.45
ω_9	47	23	2209	529	1081	22.81	0.19	520.43	0.04
ω_{10}	40	23	1600	529	920	20.00	3.00	400.08	8.99
$\sum$	400	200	16244	4096	8098	200	0.00	4039.96	56.64

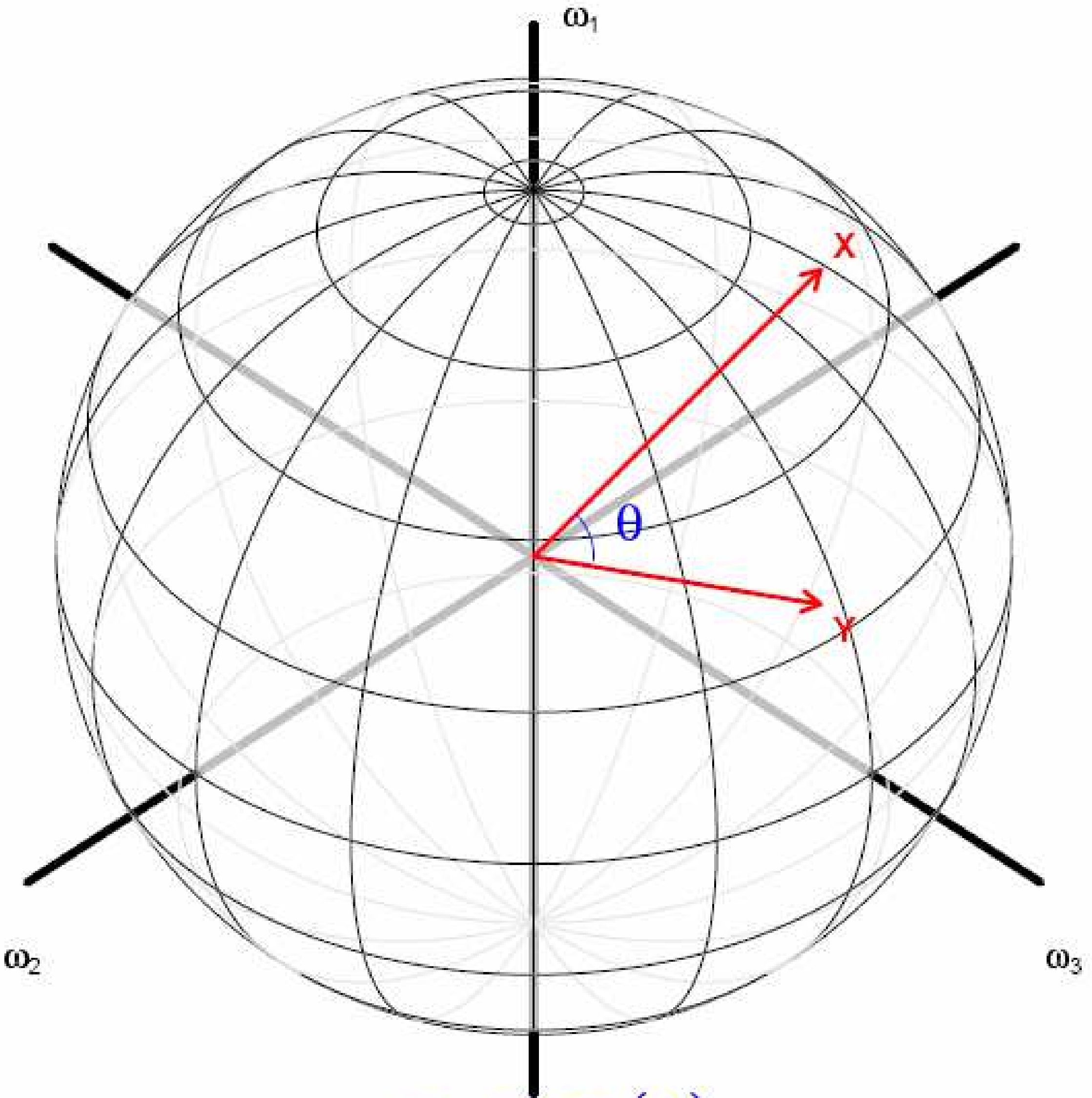
$\overline{x} = 40$	$s_x = 4.94$	$s_x^2 = 24.40$
$\overline{y} = 20$	$s_y = 3.10$	$s_y^2 = 9.60$
	$s_{xy} = 9.80$	

$$r = +0.64$$

$Y' = 0.40 \cdot X + 3.94$	$s_{y'}^2 = 3.94$
$E_Y = Y - Y'$	$s_e^2 = 5.66$

Korrelazio lineal koefizientea

Ω	X	Y	Z_x	Z_y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

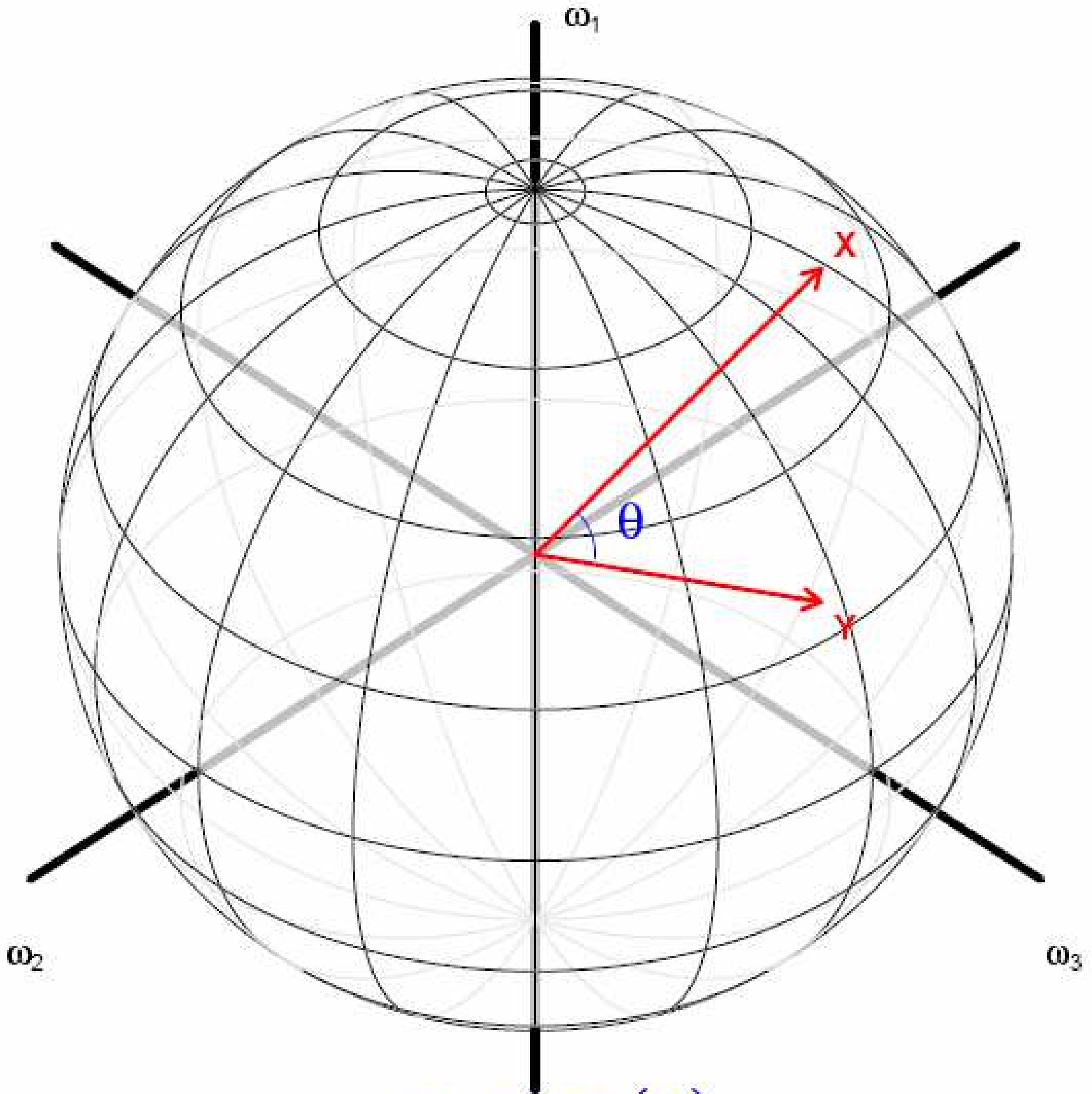


$$r = \cos(\theta)$$

Korrelazio lineal koefizientea

Ω	X	Y	Z_x	Z_y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

$$r = s_{zxzy} = \frac{1}{n}\Sigma_{i=1}^{n} z_{xi} z_{yi}$$



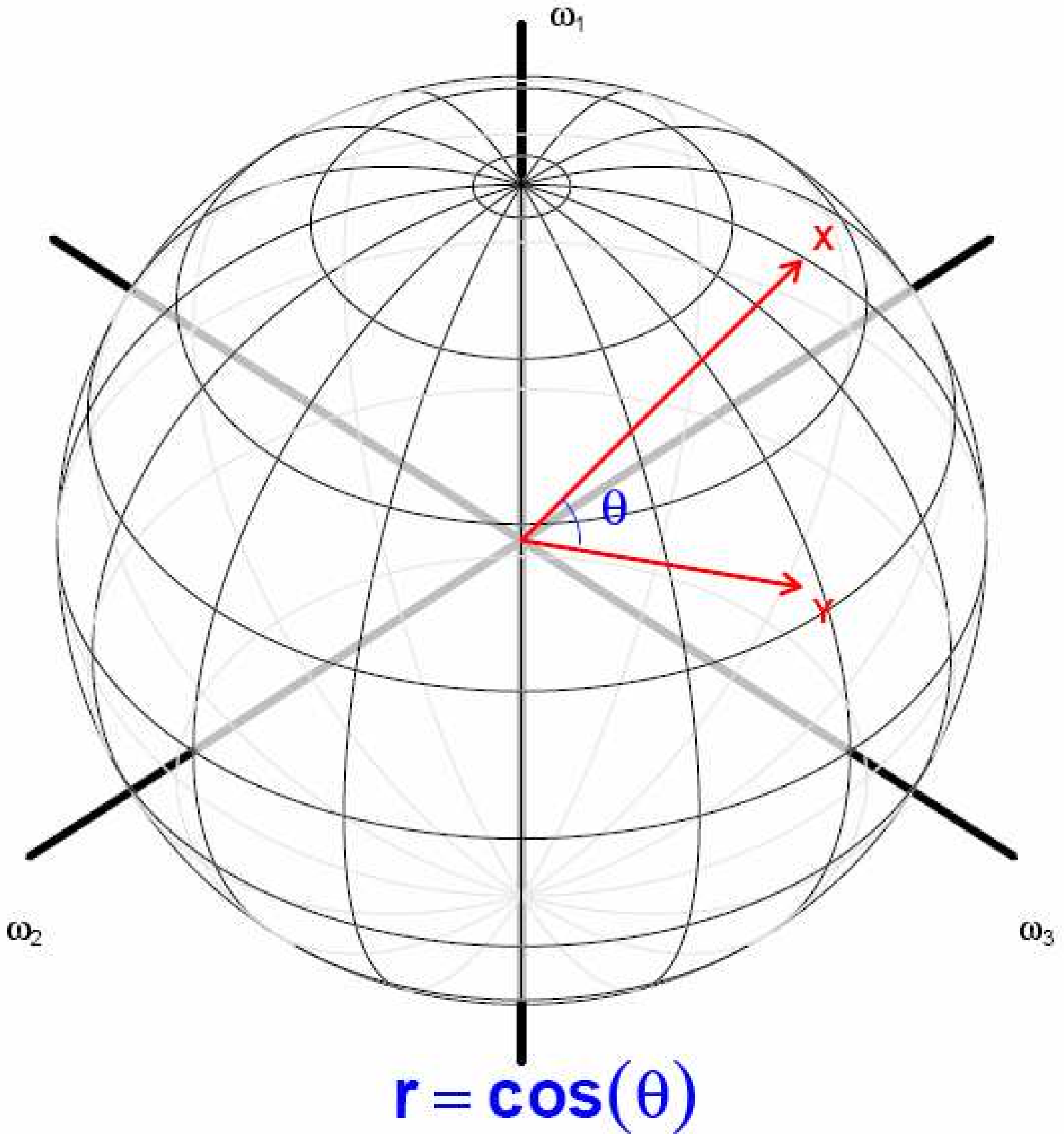
$$r = \cos(\theta)$$

Korrelazio lineal koefizientea

Ω	X	Y	Z x	Z y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

$$r = s_{zxzy} = \frac{1}{n} \Sigma_{i=1}^{n} z_{xi} z_{yi}$$

$$Z_x \bullet Z_y = \Sigma_{i=1}^{n} z_{xi} z_{yi}$$



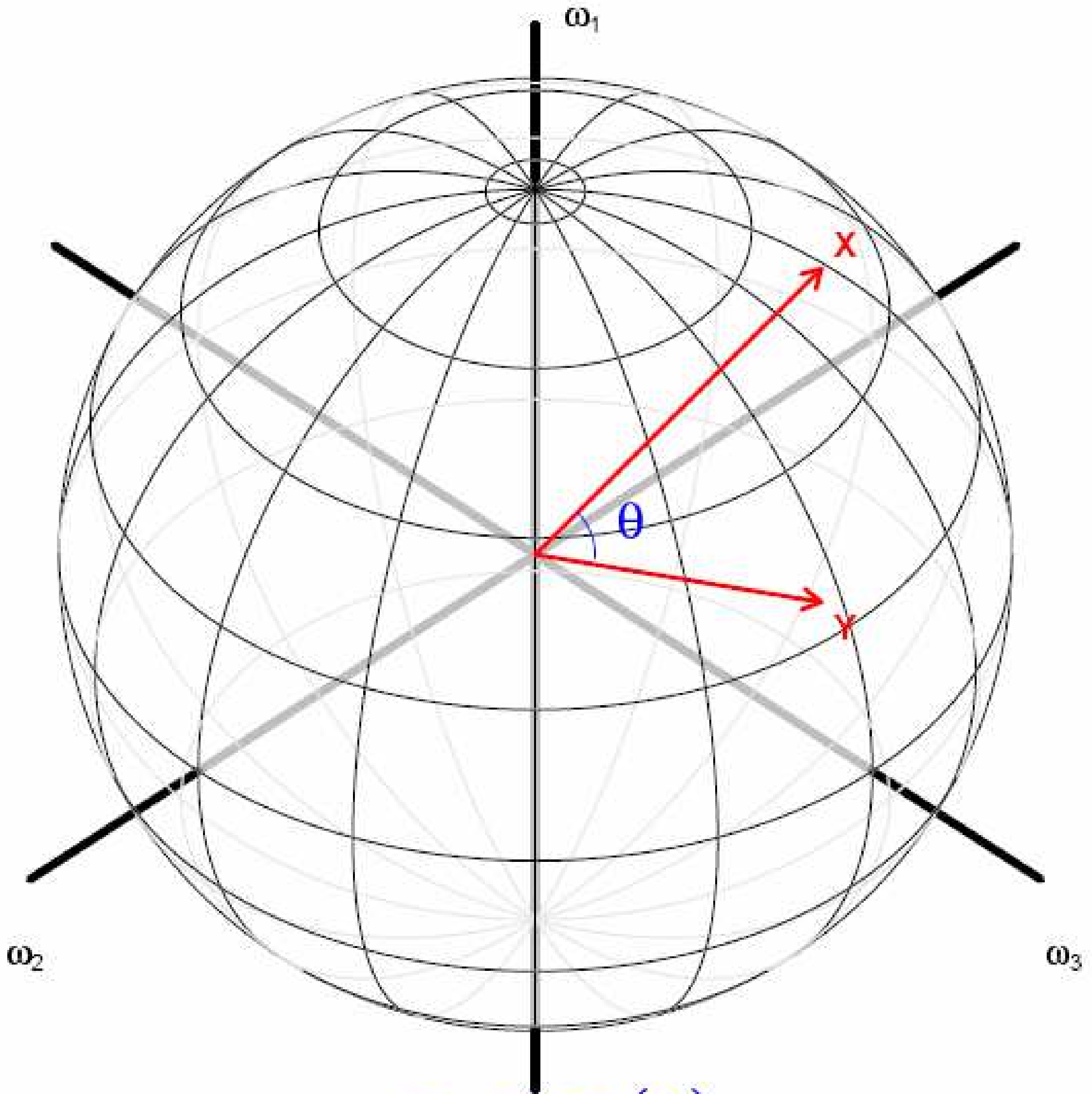
Korrelazio lineal koefizientea

Ω	X	Y	Z_x	Z_y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

$$r = s_{zxzy} = \frac{1}{n}\sum_{i=1}^{n} z_{xi} z_{yi}$$

$$Z_x \bullet Z_y = \sum_{i=1}^{n} z_{xi} z_{yi}$$

$$||Z_x||^2 = ||Z_y||^2 = n$$



$r = \cos(\theta)$

Korrelazio lineal koefizientea

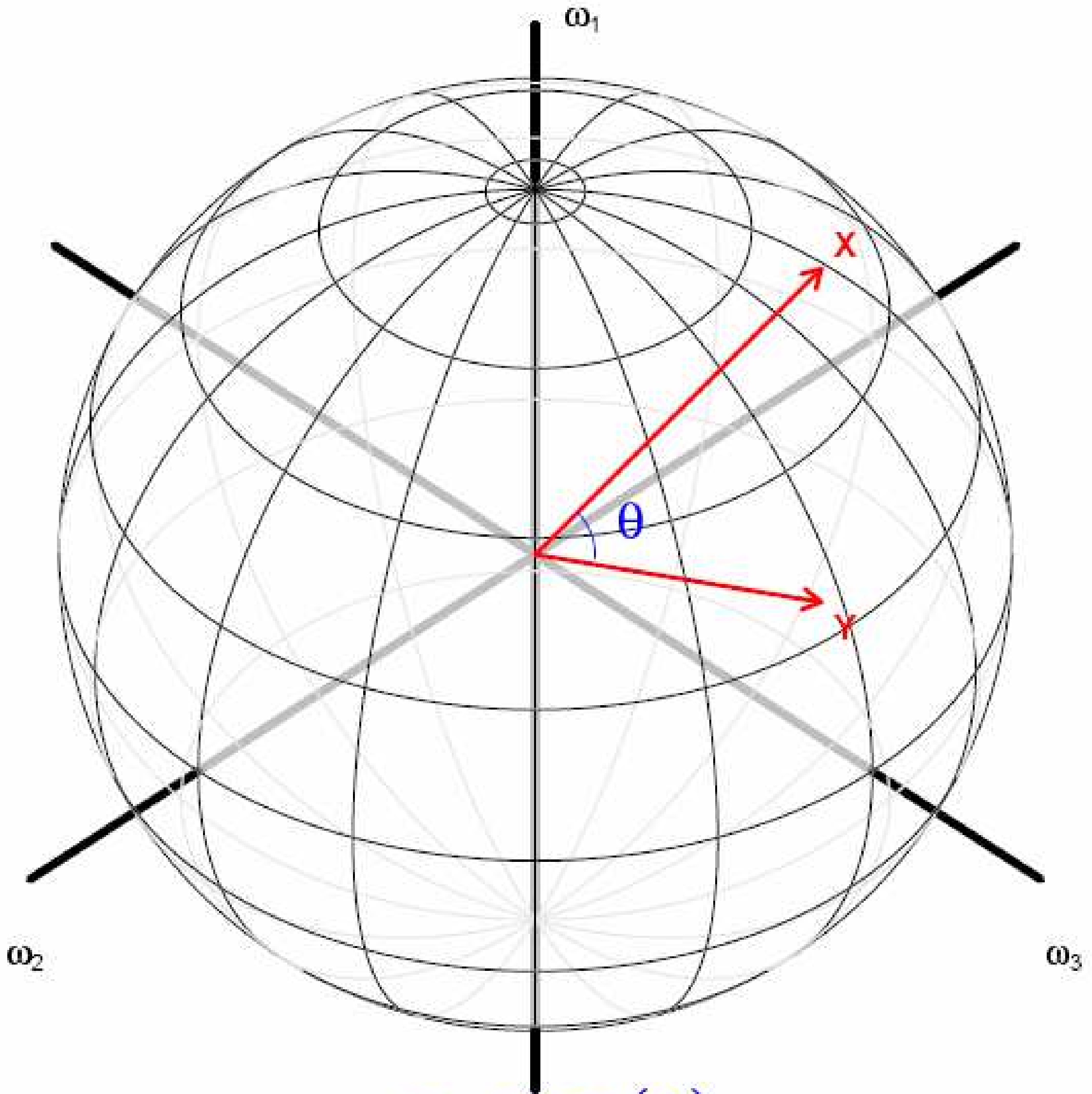
Ω	X	Y	Z_x	Z_y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

$$r = s_{zxzy} = \frac{1}{n}\Sigma_{i=1}^{n} z_{xi} z_{yi}$$

$$Z_x \bullet Z_y = \Sigma_{i=1}^{n} z_{xi} z_{yi}$$

$$||Z_x||^2 = ||Z_y||^2 = n$$

$$Z_x \bullet Z_y = n \cdot cos(\theta)$$



$$r = \cos(\theta)$$

Korrelazio lineal koefizientea

Ω	X	Y	Z x	Z y
ω_1	x_1	y_1	z_{x1}	z_{y1}
ω_i	x_i	y_i	z_{xi}	z_{yi}
ω_n	x_n	y_n	z_{xn}	z_{yn}

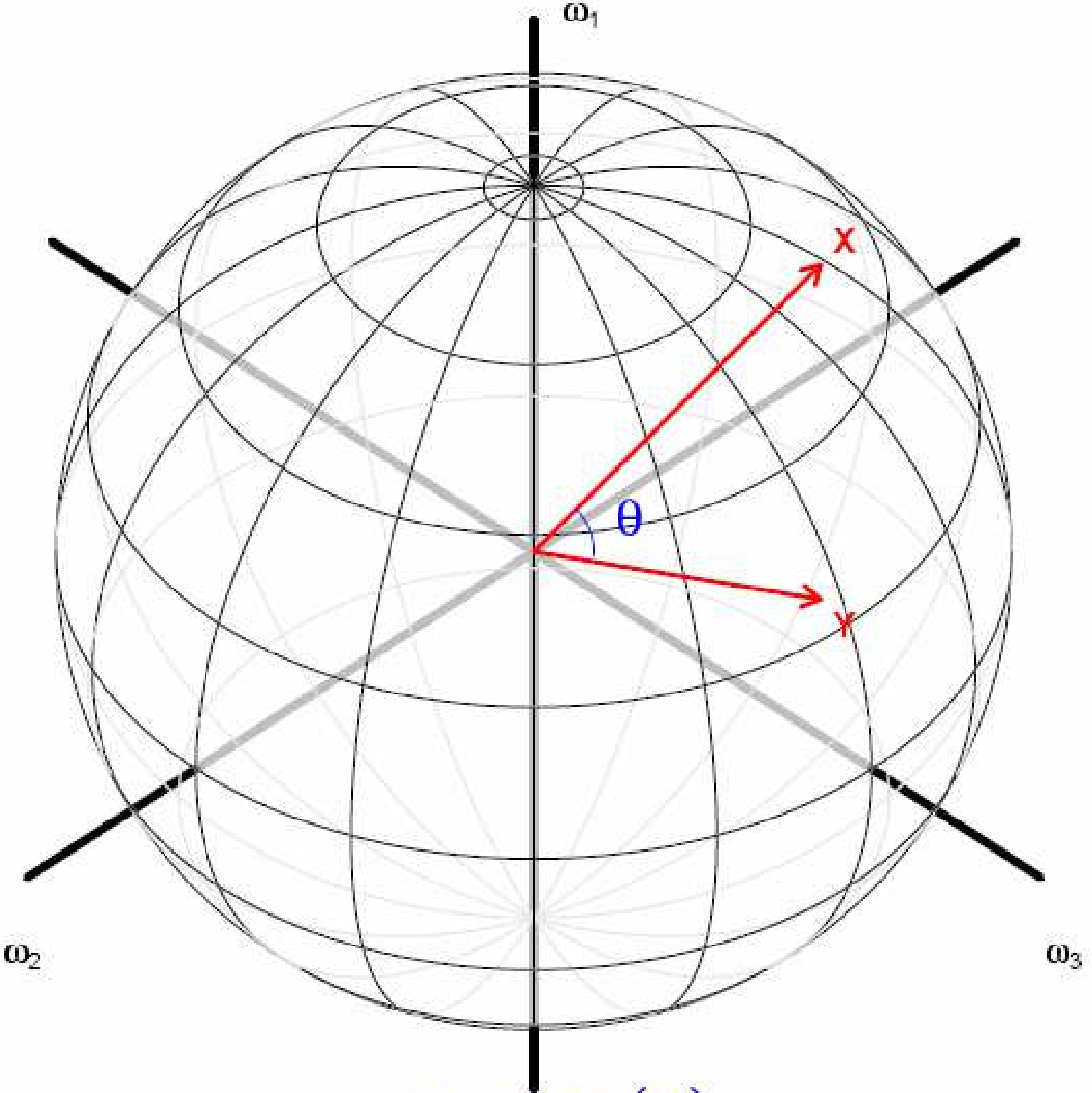
$$r = s_{zxzy} = \frac{1}{n}\Sigma_{i=1}^{n} z_{xi} z_{yi}$$

$$Z_x \bullet Z_y = \Sigma_{i=1}^{n} z_{xi} z_{yi}$$

$$||Z_x||^2 = ||Z_y||^2 = n$$

$$Z_x \bullet Z_y = n \cdot cos(\theta)$$

$$r = \frac{1}{n} Z_x \bullet Z_y = cos(\theta)$$



$r = \cos(\theta)$

Korrelazio lineal koefizientea

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

$$-1 \leq r \leq +1$$

$$\frac{Y - \overline{y}}{s_y} = r \cdot \frac{X - \overline{x}}{s_x} + E^*$$

$$r = +1 \qquad \frac{X(\omega_k) - \overline{x}}{s_x} = \frac{Y(\omega_k) - \overline{y}}{s_y} \qquad Y = \frac{s_y}{s_x} \cdot X + \left[\overline{y} - \frac{s_y}{s_x} \cdot \overline{x}\right]$$

$$r = -1 \qquad \frac{X(\omega_k) - \overline{x}}{s_x} = -\frac{Y(\omega_k) - \overline{y}}{s_y} \qquad Y = -\frac{s_y}{s_x} \cdot X + \left[\overline{y} + \frac{s_y}{s_x} \cdot \overline{x}\right]$$

$r = 0$ *X* eta *Y* *inkorrelatuak* $Y' = \overline{y}$

- Baldin badira *X* eta *Y* elkar askeak, *orduan* $r = 0$

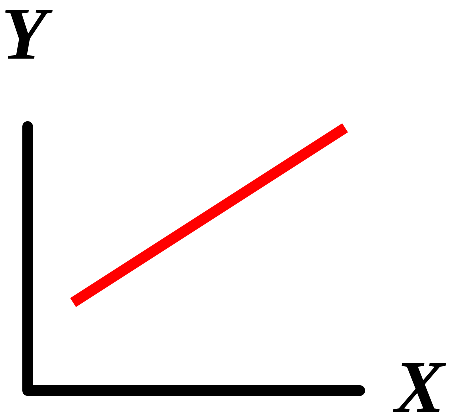
- *X* eta *Y* elkar askeak *baldin eta soilik baldin* $r = 0$ eta (*X*,*Y*) bi dimentsioko normala

*X*rekiko *Y*ren erregresio lineal bakuna

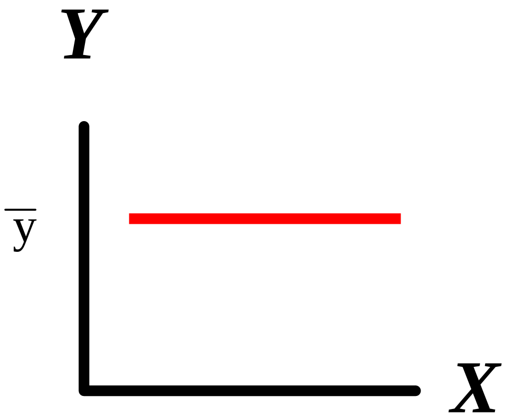
$$Y' = b + aX$$

$$a = \frac{s_{xy}}{s_x^{\ 2}} \qquad\qquad r = \frac{s_{xy}}{s_x\, s_y}$$

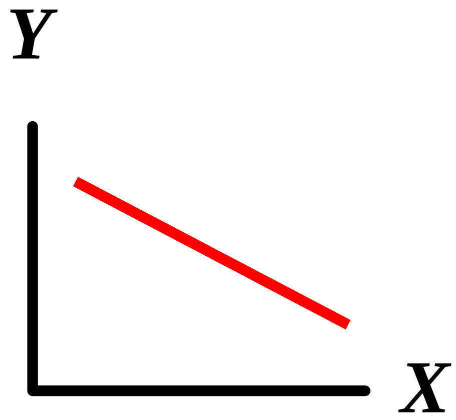
$s_{xy} > 0$
$a > 0$
$r > 0$



$s_{xy} = 0$
$a = 0$
$r = 0$



$s_{xy} < 0$
$a < 0$
$r < 0$



Korrelazio lineal koefizientea

Kalkulatu aldagai hauen arteko binakako korrelazio lineal koefizientearen balioak

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

Korrelazio lineal koefizientea

Kalkulatu aldagai hauen arteko binakako korrelazio lineal koefizientearen balioak.

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

http://www.r-project.org/ RWeka

http://www.cs.waikato.ac.nz/ml/weka/

WEKA The University of Waikato

Korrelazio lineal koefizientea

Kalkulatu aldagai hauen arteko binakako korrelazio lineal koefizientearen balioak.

Kostua	(Y)	22.6	15.0	78.1	28.0	80.5	24.5	20.5	147.6	4.2	48.2	20.5
Fitxategiak	(X_1)	4	2	20	6	6	3	4	16	4	6	5
Fluxuak	(X_2)	44	33	80	24	227	20	41	187	19	50	48
Prozesuak	(X_3)	18	15	80	21	50	18	13	137	15	21	17

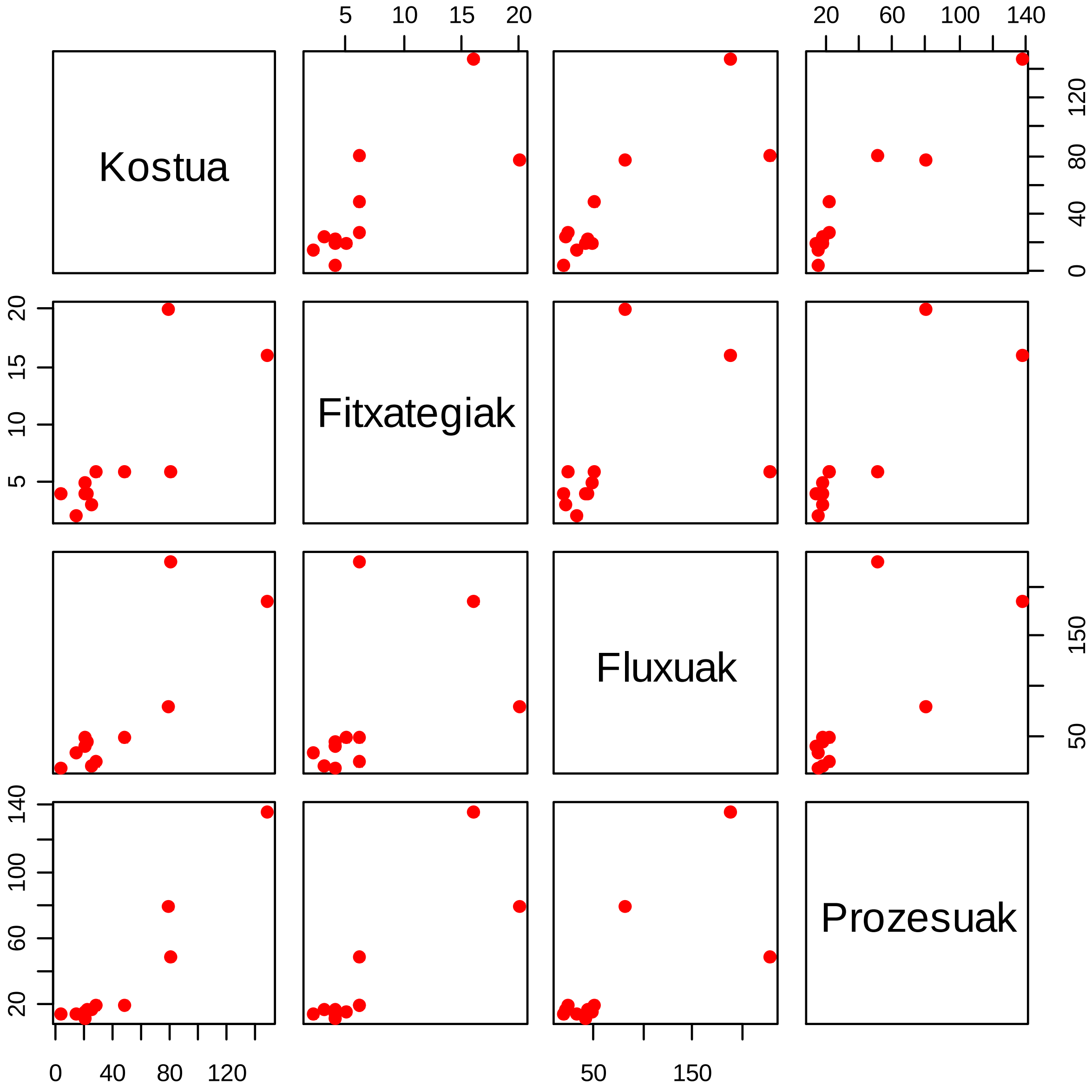
http://www.r-project.org/ RWeka

http://www.cs.waikato.ac.nz/ml/weka/

```
> Kostua <- c(22.6,15.0,78.1,28.0,80.5,24.5,20.5,147.6,4.2,48.2,20.5)
> Fitxategiak <- c(4,2,20,6,6,3,4,16,4,6,5)
> Fluxuak <- c(44,33,80,24,227,20,41,187,19,50,48)
> Prozesuak <- c(18,15,80,21,50,18,13,137,15,21,17)
> datuak <- data.frame(Kostua,Fitxategiak,Fluxuak,Prozesuak)
> attach(datuak)
> cor(datuak)
>                Kostua Fitxategiak   Fluxuak Prozesuak
> Kostua      1.0000000   0.7784743 0.8303919 0.9598421
> Fitxategiak 0.7784743   1.0000000 0.4589818 0.8545609
> Fluxuak     0.8303919   0.4589818 1.0000000 0.7204369
> Prozesuak   0.9598421   0.8545609 0.7204369 1.0000000
```

Korrelazio lineal koefizientea

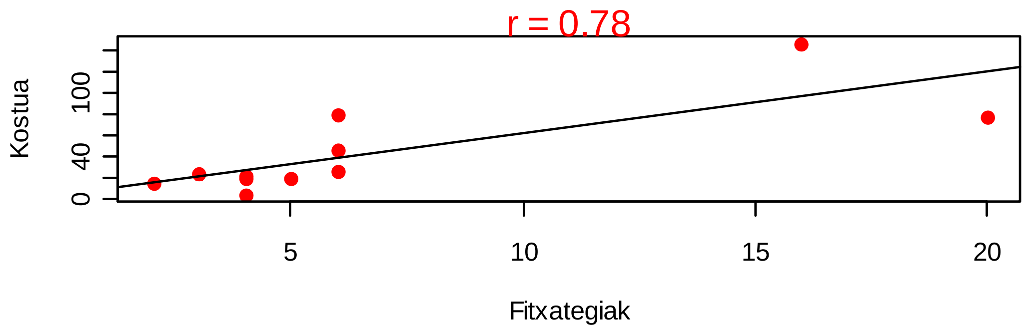
```
> pairs(datuak, pch=19, col="red")
```

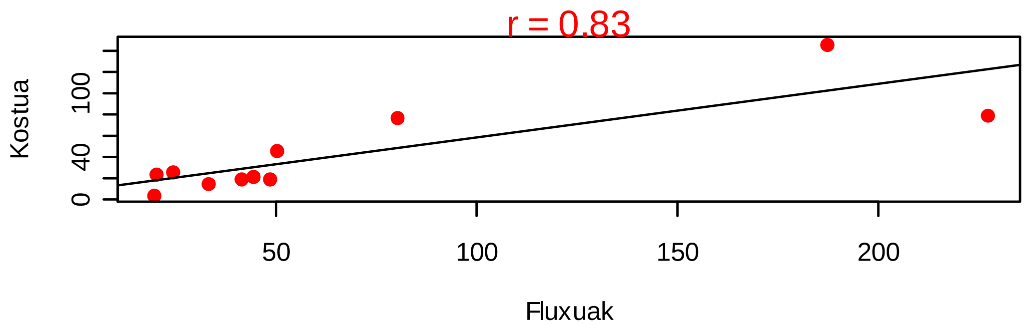


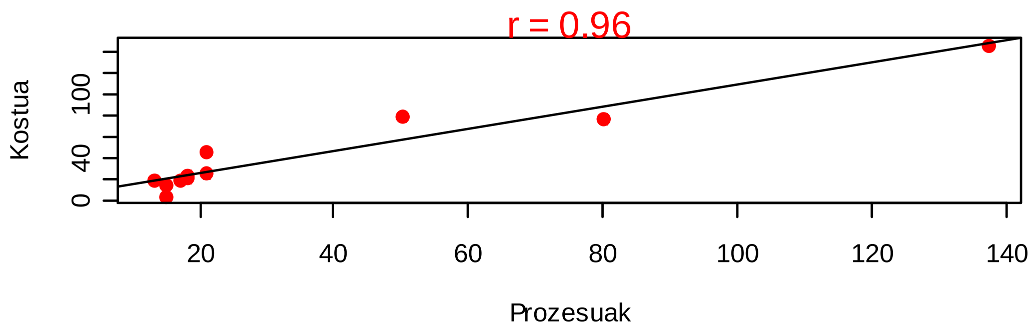
Korrelazio lineal koefizientea

```
> par(mfrow=c(3,1))
> plot(Fitxategiak, Kostua,  pch=19, col="red")
> regKostuaFitx <- lm(Kostua~Fitxategiak)
> abline(a=regKostuaFitx[1], b=regKostuaFitx[2] )
> mtext(eval(expression(paste("r =",
+ round(cor(Kostua,Fitxategiak),digits=2)))), 3, col="red")
> plot(Fluxuak, Kostua,  pch=19, col="red")
> regKostuaFlux <- lm(Kostua~Fluxuak)
> abline(a=regKostuaFlux[1], b=regKostuaFlux[2] )
> mtext(eval(expression(paste("r =",
+ round(cor(Kostua,Fluxuak),digits=2)))), 3, col="red")
> plot(Prozesuak, Kostua, pch=19, col="red")
> regKostuaProz <- lm(Kostua~Prozesuak)
> abline(a=regKostuaProz[1], b=regKostuaProz[2] )
> mtext(eval(expression(paste("r =",
+ round(cor(Kostua,Prozesuak),digits=2)))), 3, col="red")
>
```

Korrelazio lineal koefizientea

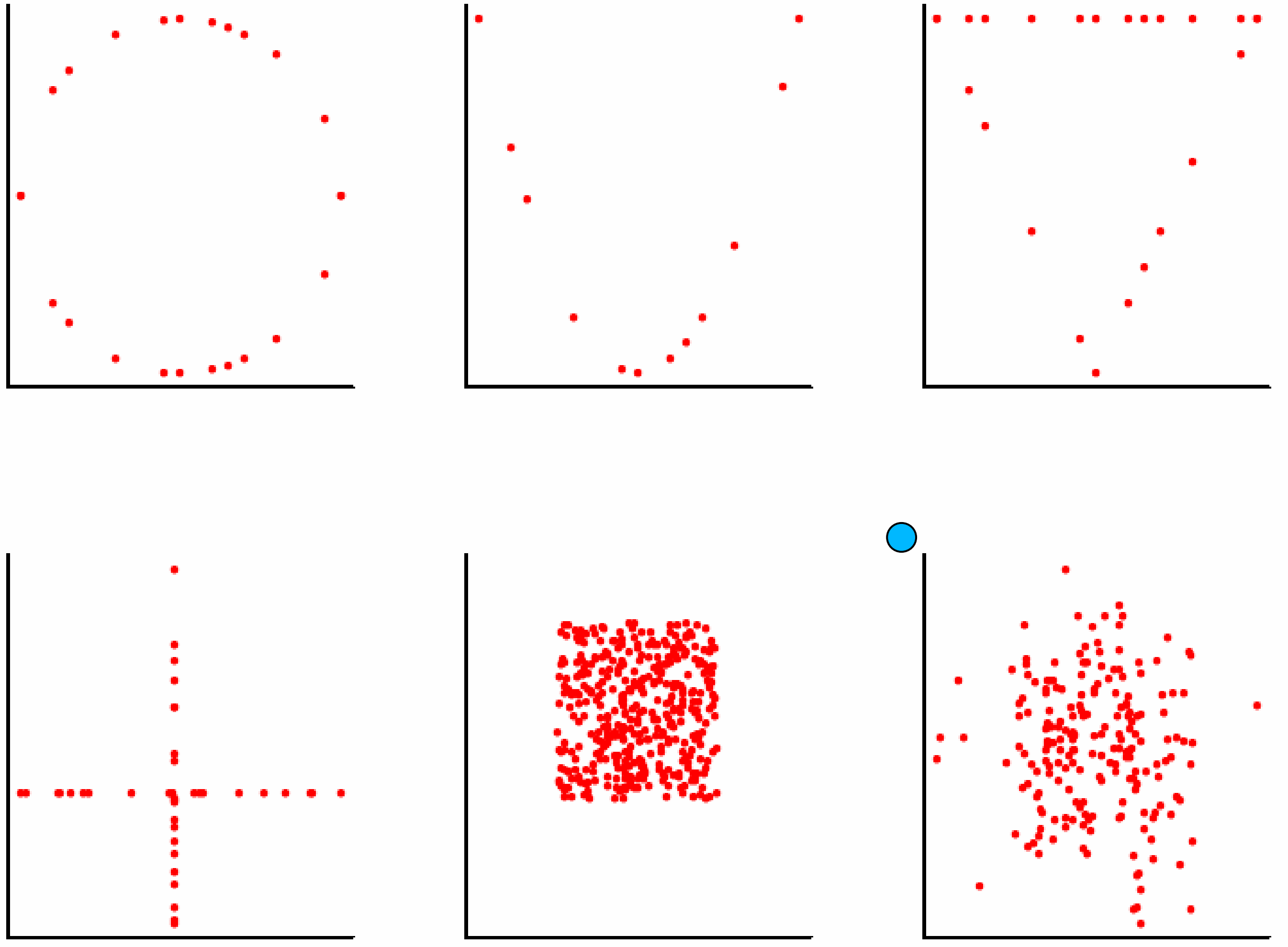






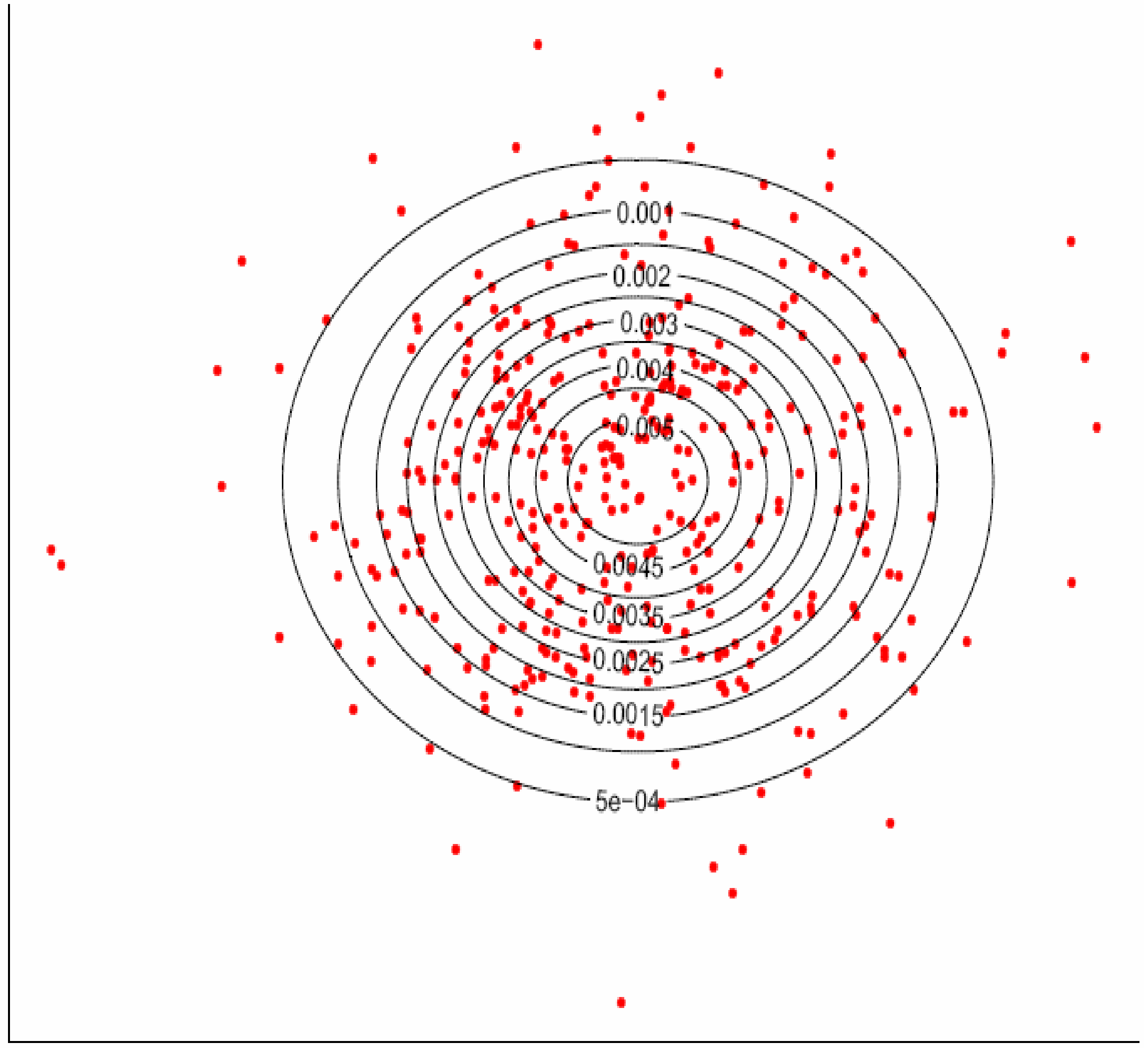
Korrelazio lineal koefizientea

r=0 Bi aldagaiak linealki *inkorrelatuak* daude, baina *ez* dira *elkar askeak*, azken kasua izan ezik



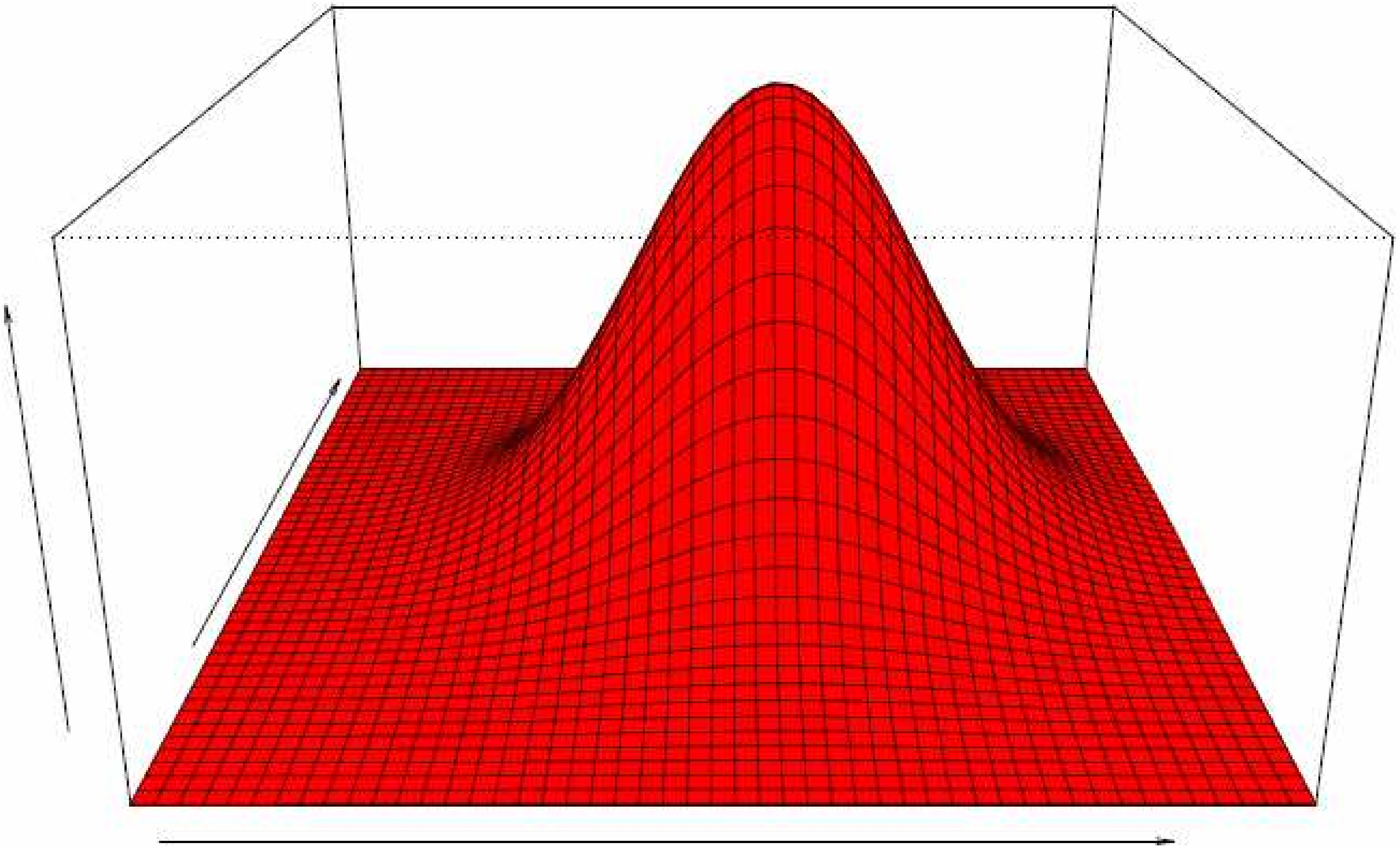
Korrelazio lineal koefizientea

r=0 Bi aldagaiak linealki *inkorrelatuak* daude, eta *elkar askeak* dira



Korrelazio lineal koefizientea

ρ=0 Bi aldagaiak *elkar askeak* dira



$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \; e^{-\frac{1}{2}\frac{1}{1-\rho^2}[(\frac{x-\mu_x}{\sigma_x})^2-2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})+(\frac{y-\mu_y}{\sigma_y})^2]}$$

Korrelazio lineal koefizientea

Scatterplots of Anscombe's Quartet of Datasets

x1	y1		x2	y2		x3	y3		x4	y4
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.1		14	8.84		8	7.04
6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.1		4	5.39		19	12.5
12	10.84		12	9.13		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

Korrelazio lineal koefizientea

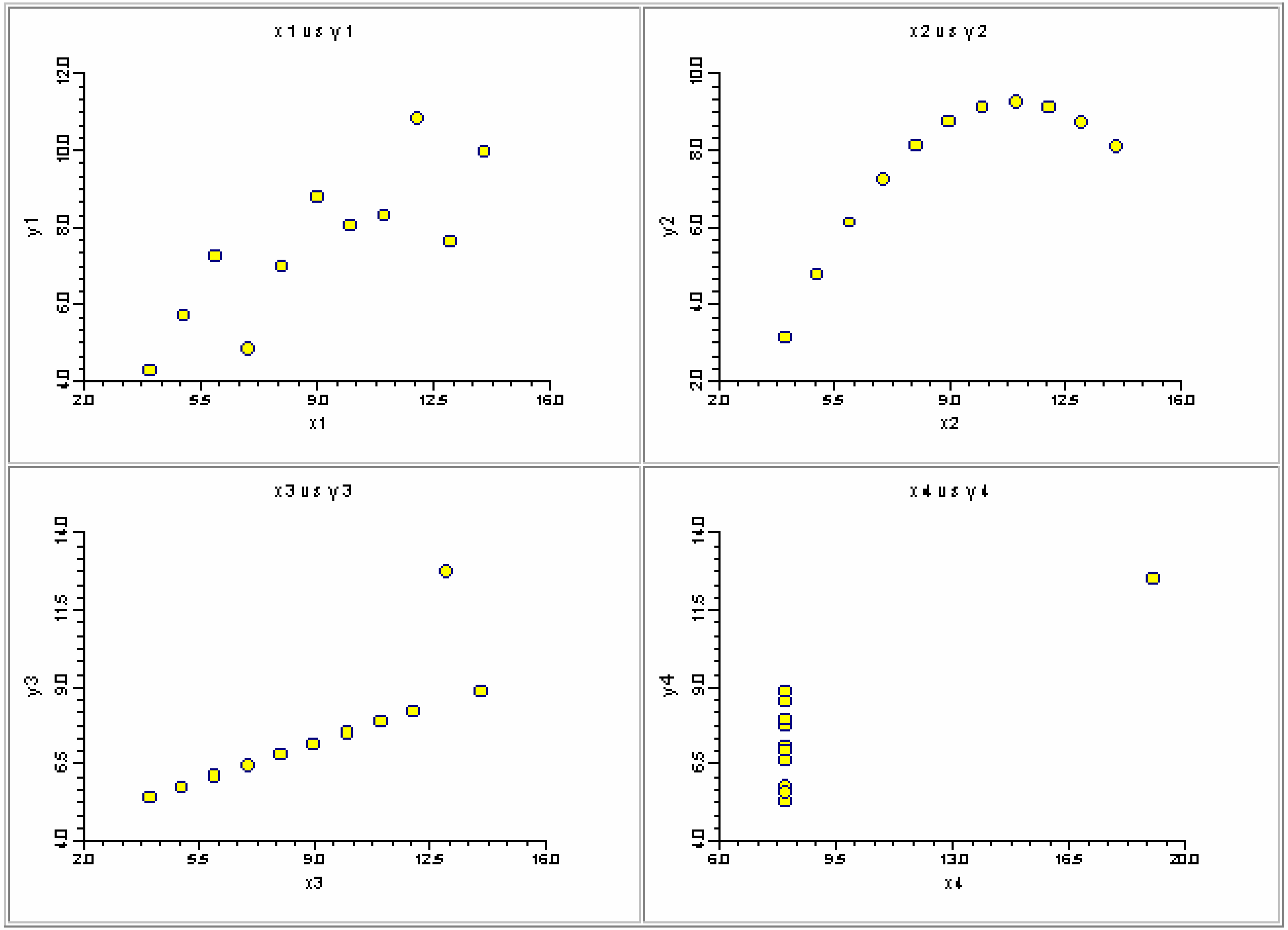
Scatterplots of Anscombe's Quartet of Datasets

Each dataset has the following identical set of summary statistics:

N	11
Mean of x's	9.0
Mean of y's	7.5
Equation of regression line	y = 3 + 0.5x
correlation coefficient	0.82

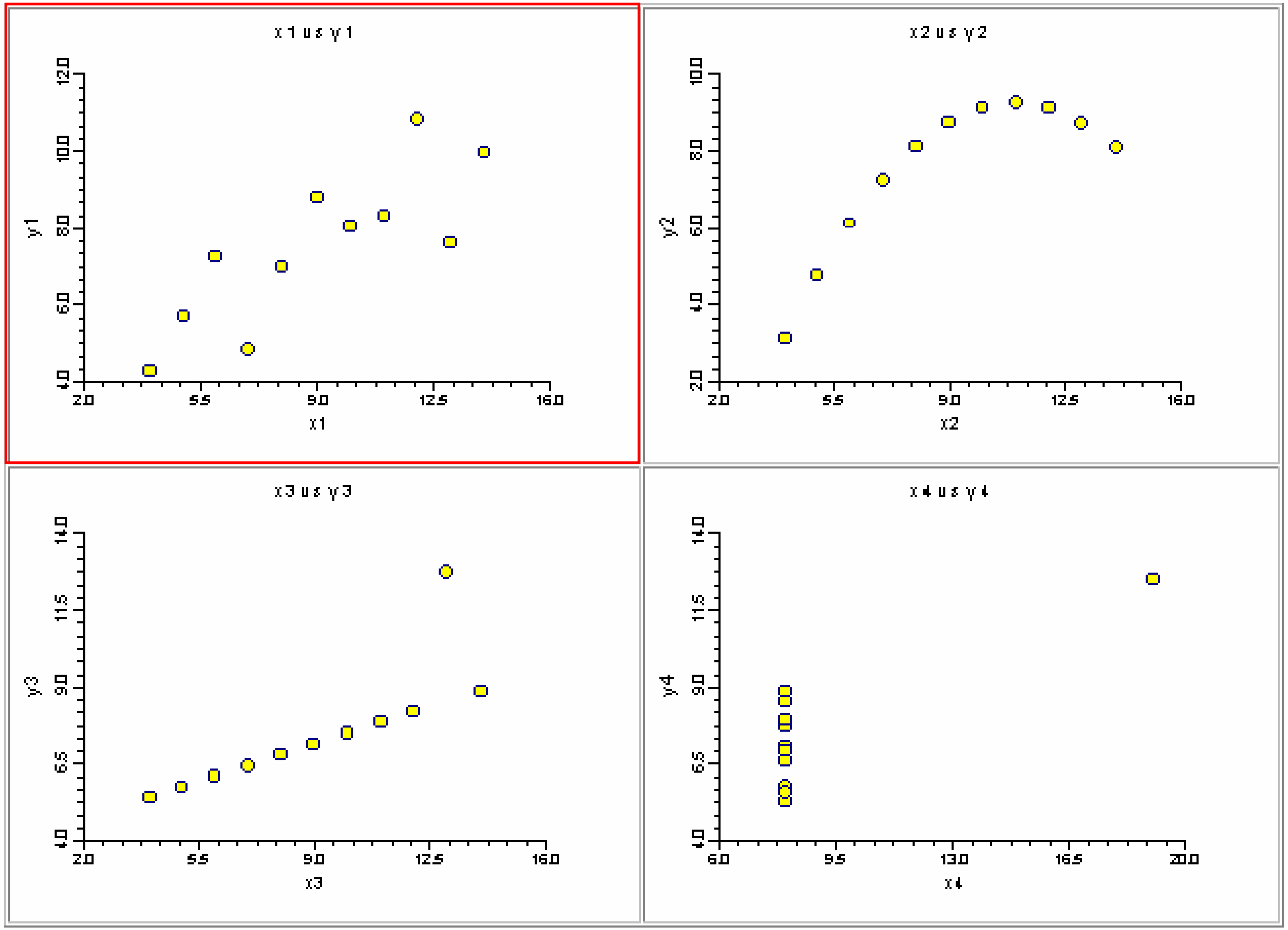
Korrelazio lineal koefizientea

Scatterplots of Anscombe's Quartet of Datasets   r=0.82



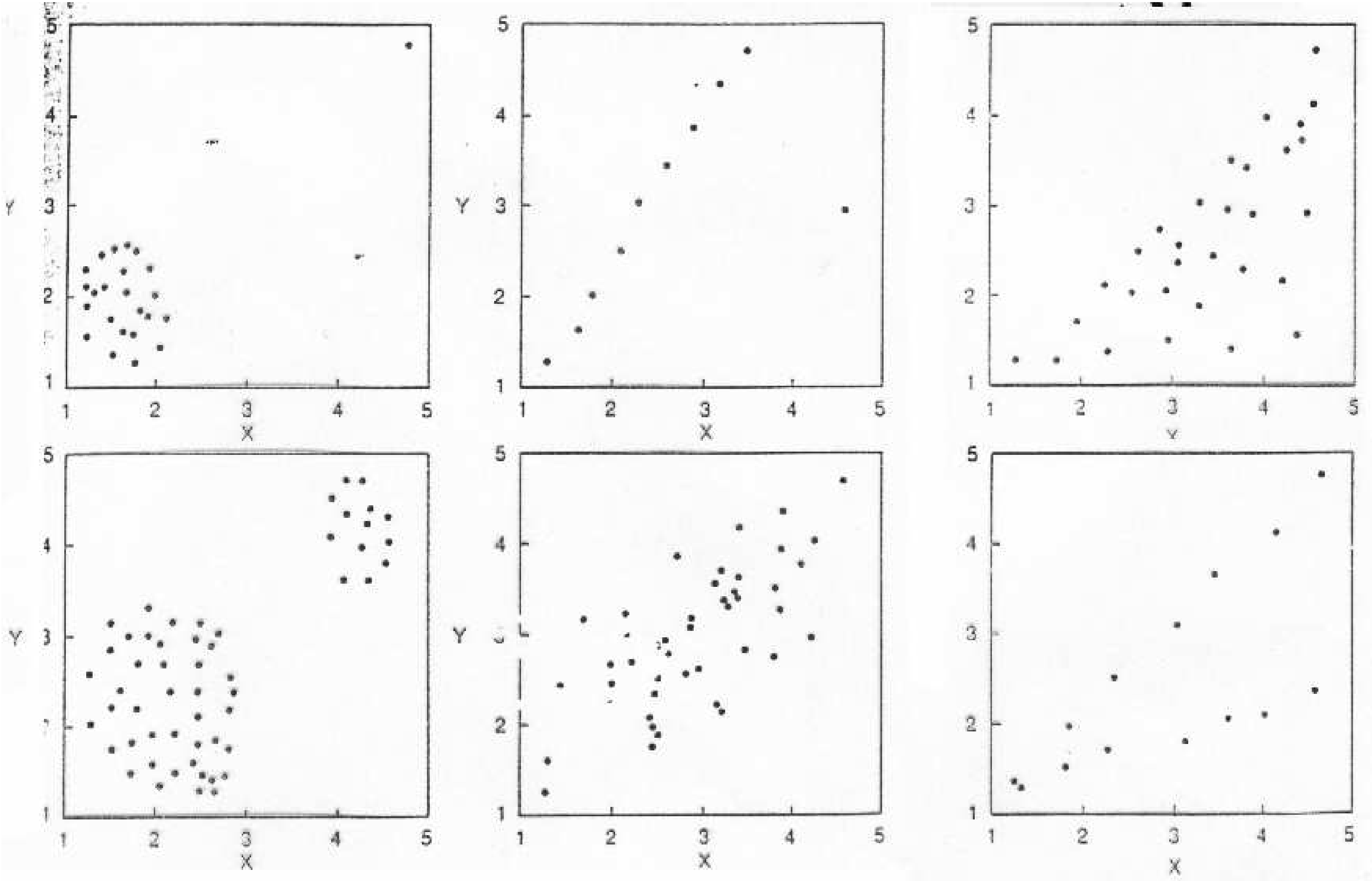
Korrelazio lineal koefizientea

Scatterplots of Anscombe's Quartet of Datasets **r=0.82**



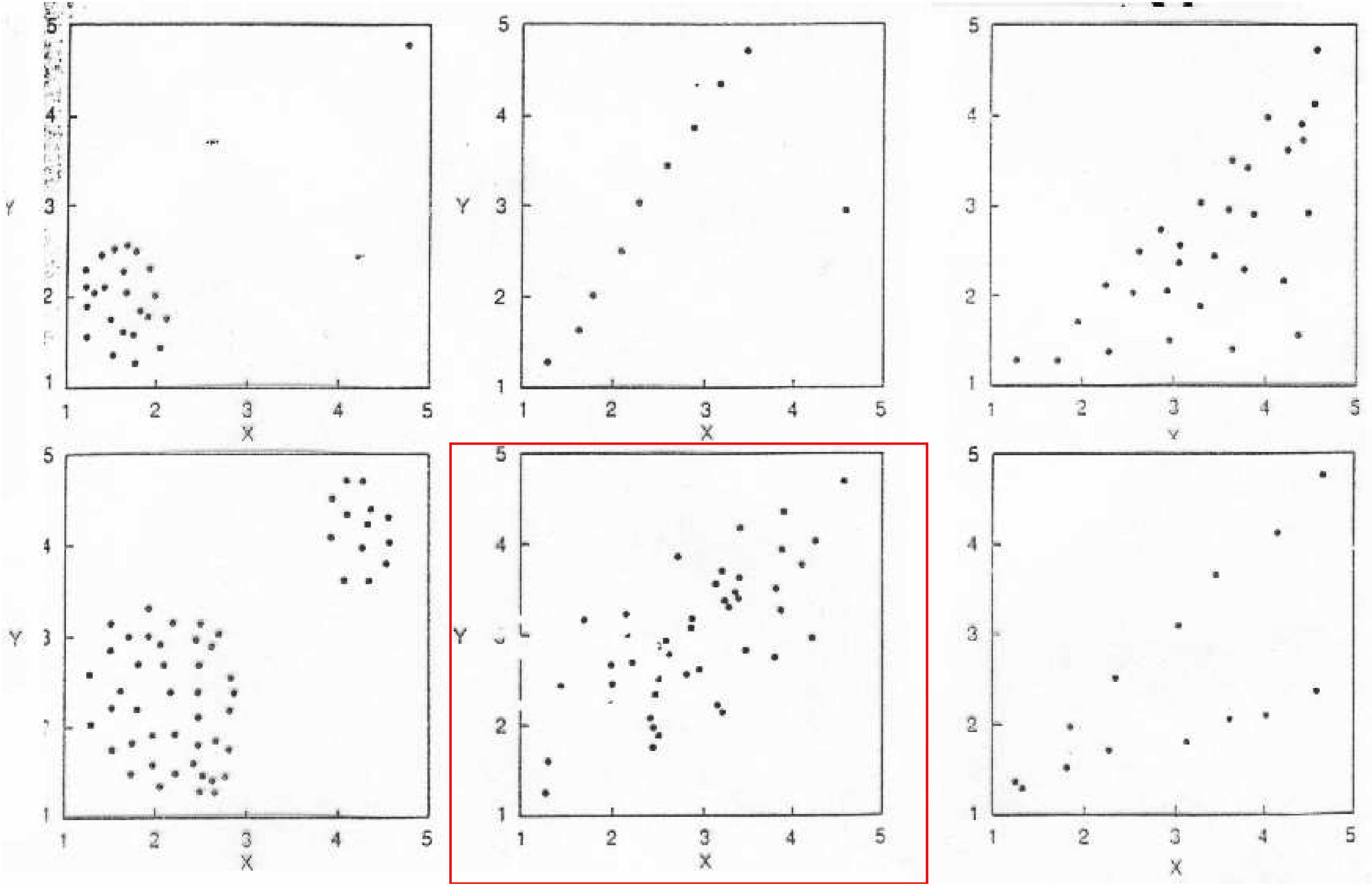
Korrelazio lineal koefizientea

r=0.7

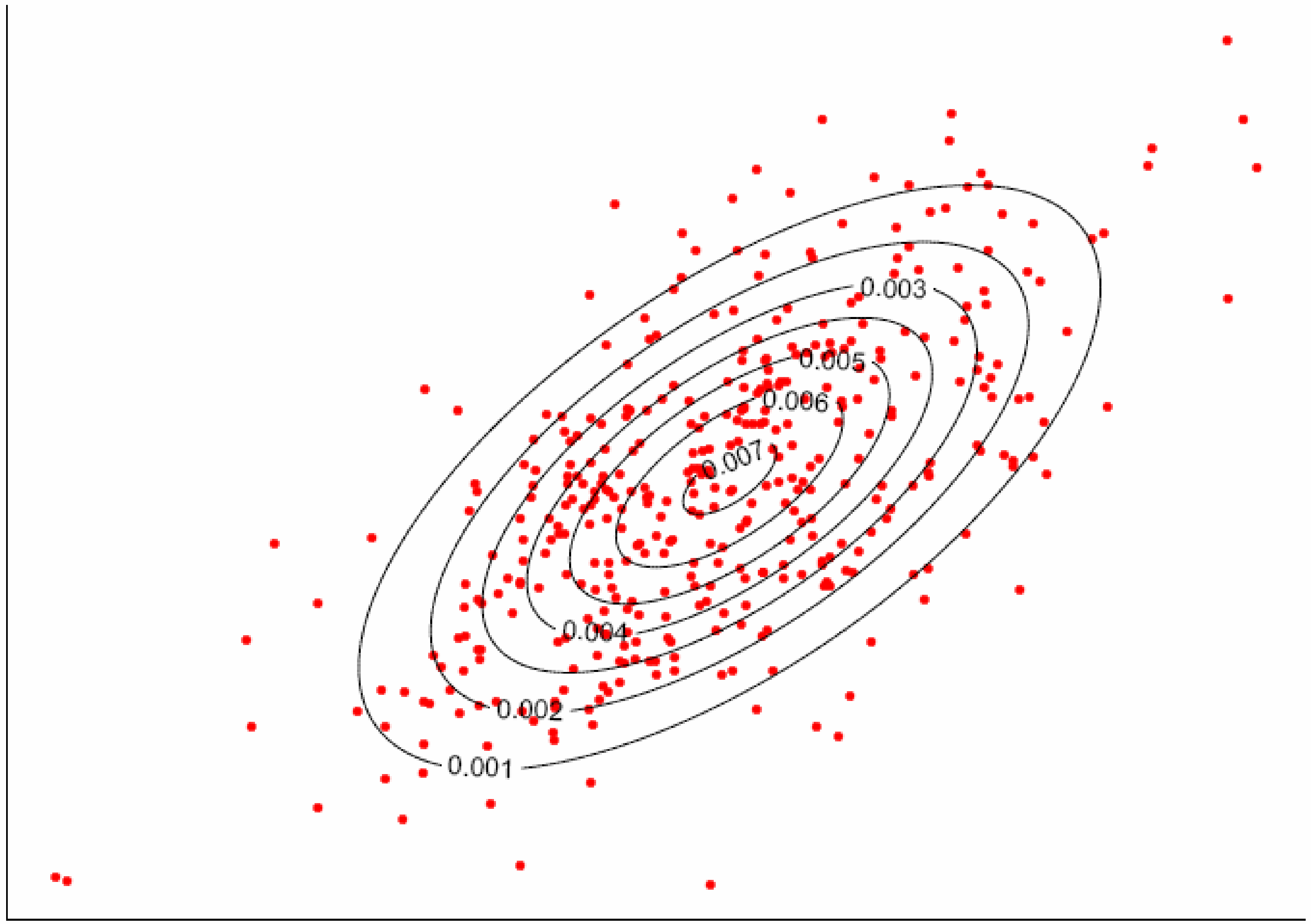


Korrelazio lineal koefizientea

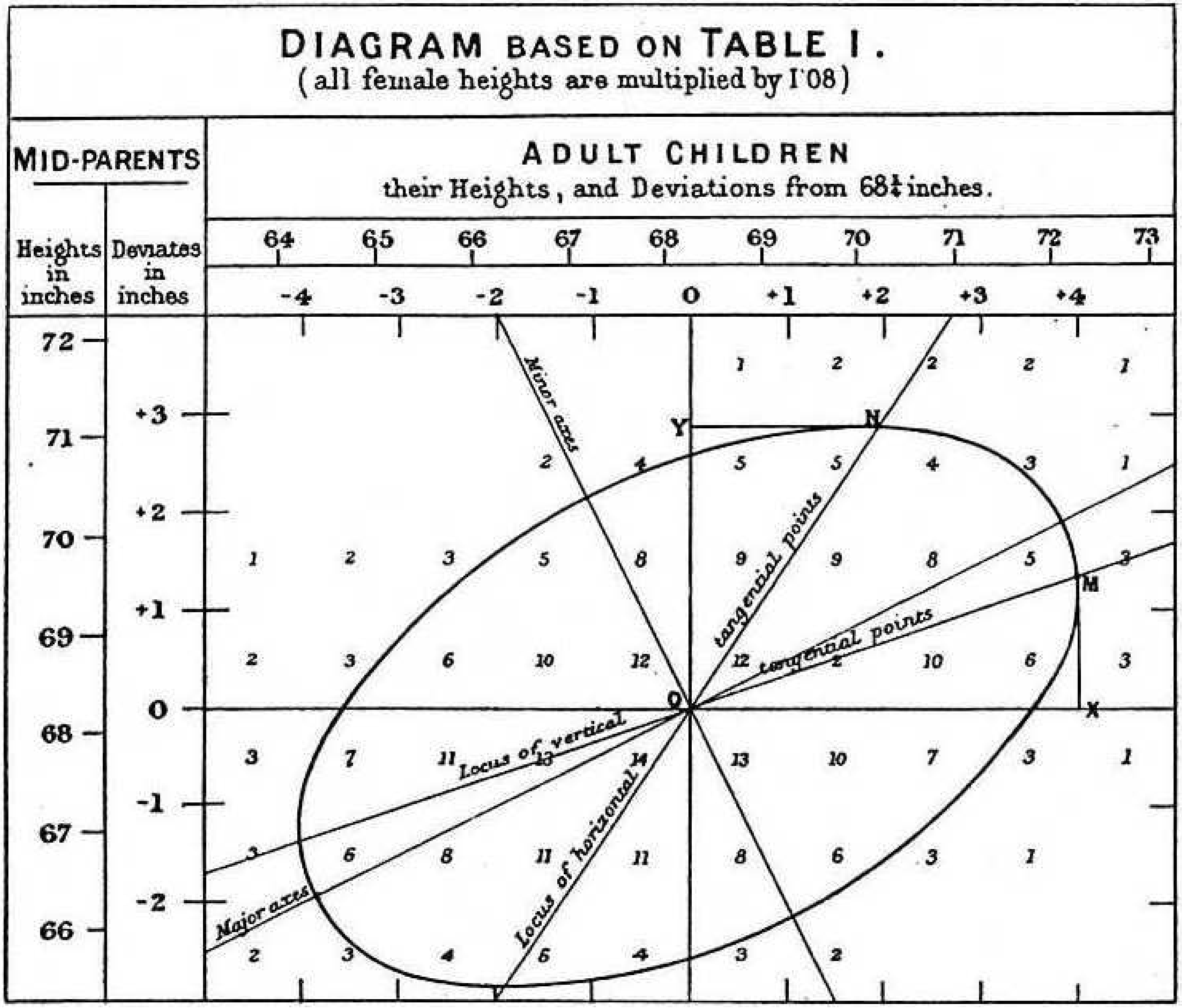
r=0.7



Korrelazio lineal koefizientea

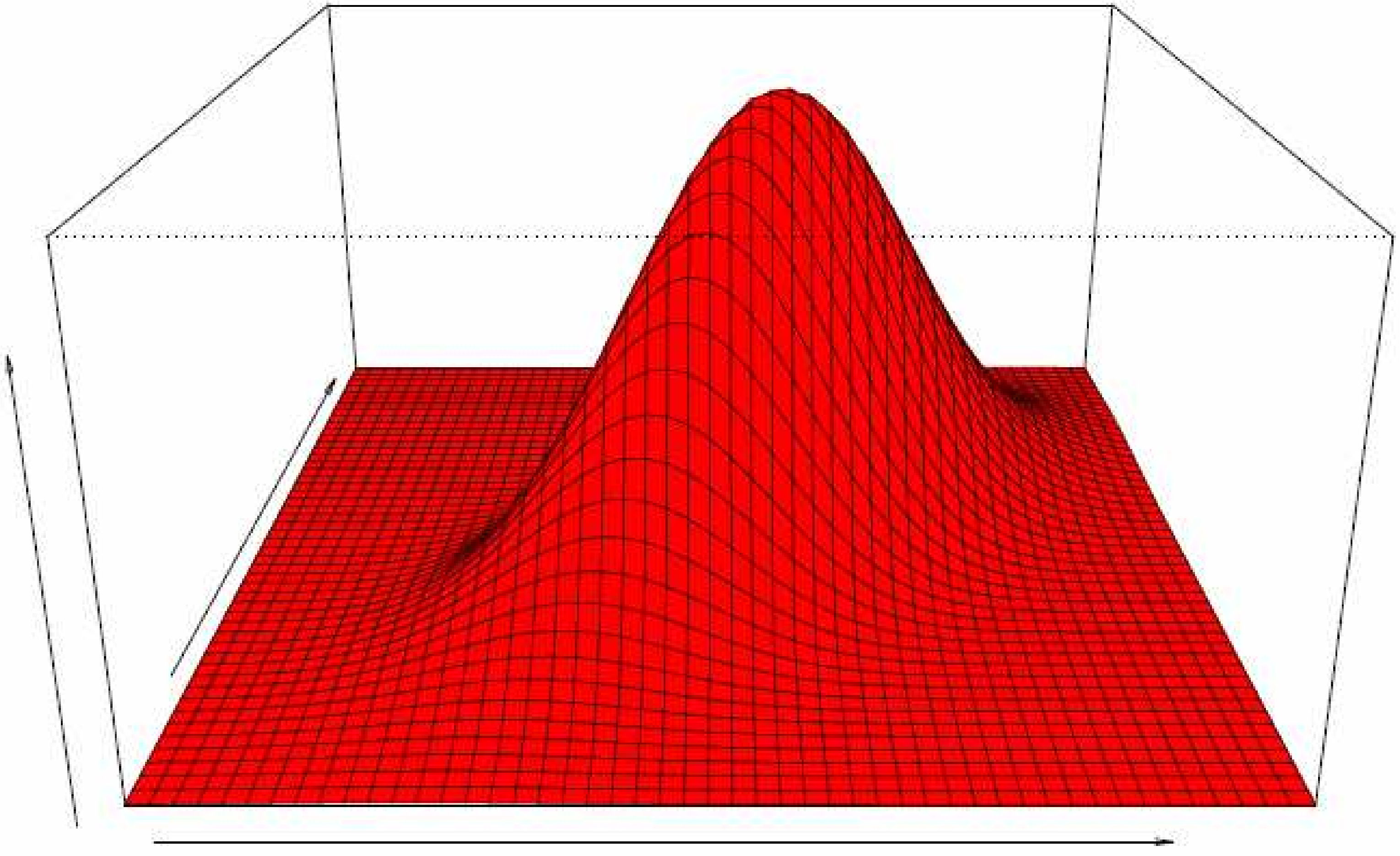


Korrelazio lineal koefizientea



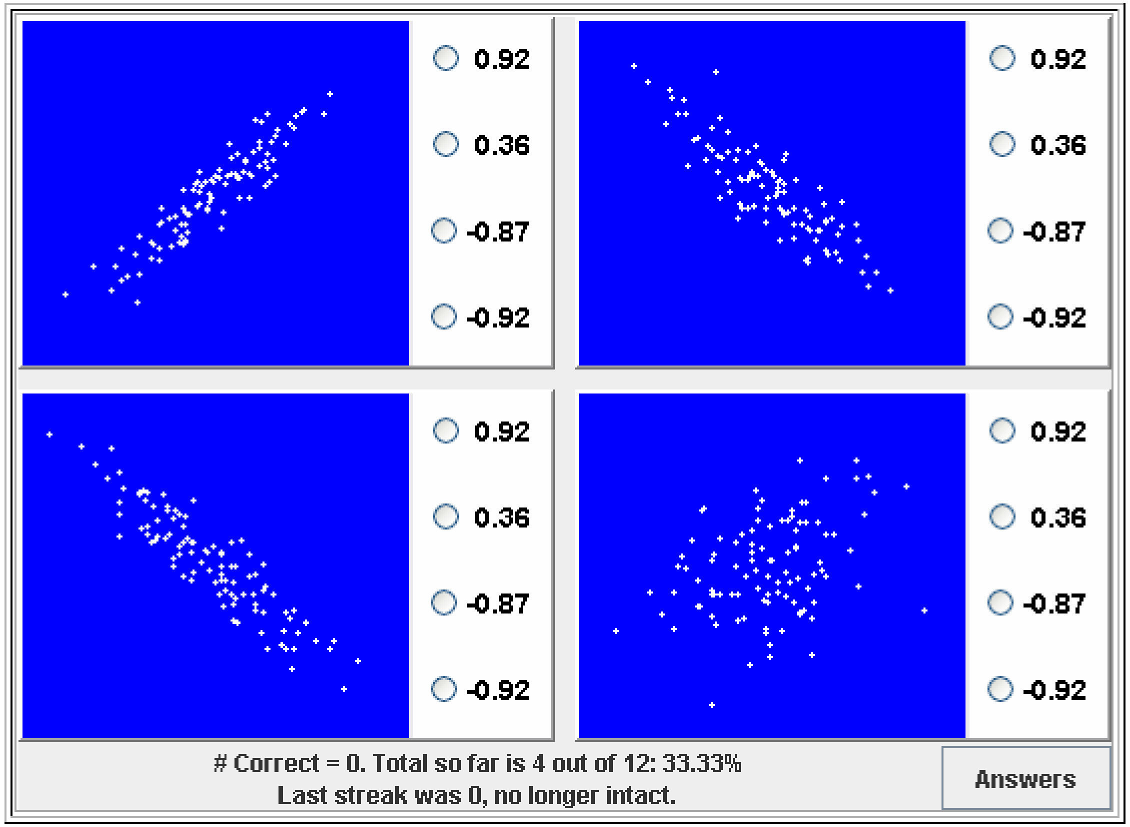
Korrelazio lineal koefizientea

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \; e^{-\frac{1}{2}\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]}$$



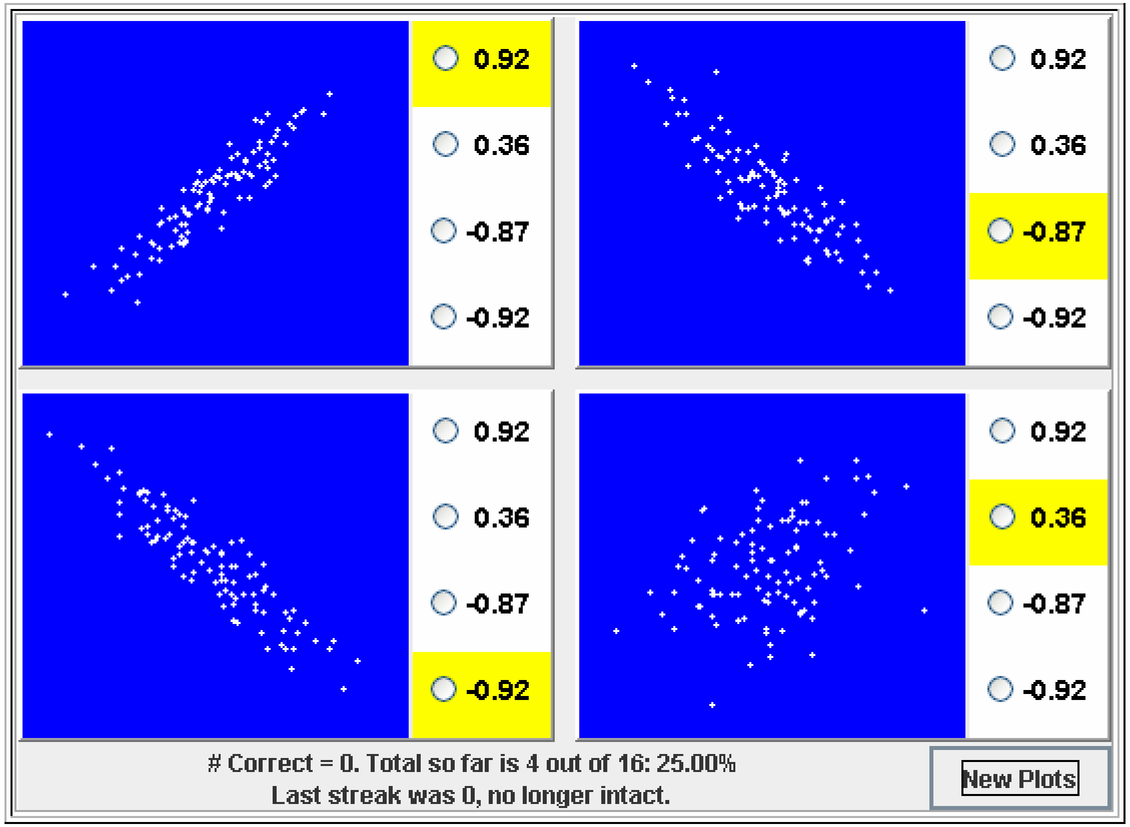
Korrelazio lineal koefizientea

http://istics.net/stat/Correlations/



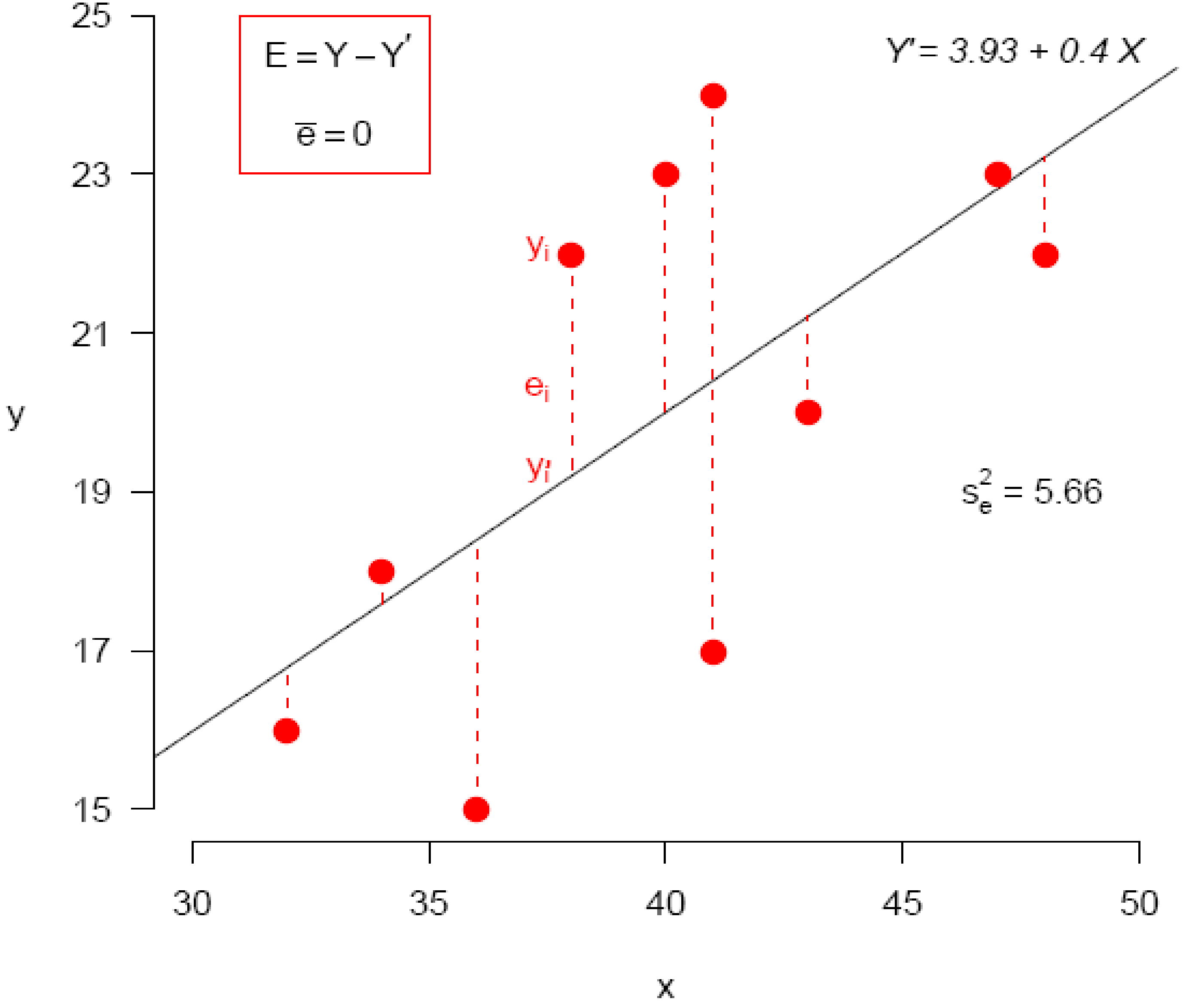
Korrelazio lineal koefizientea

http://istics.net/stat/Correlations/

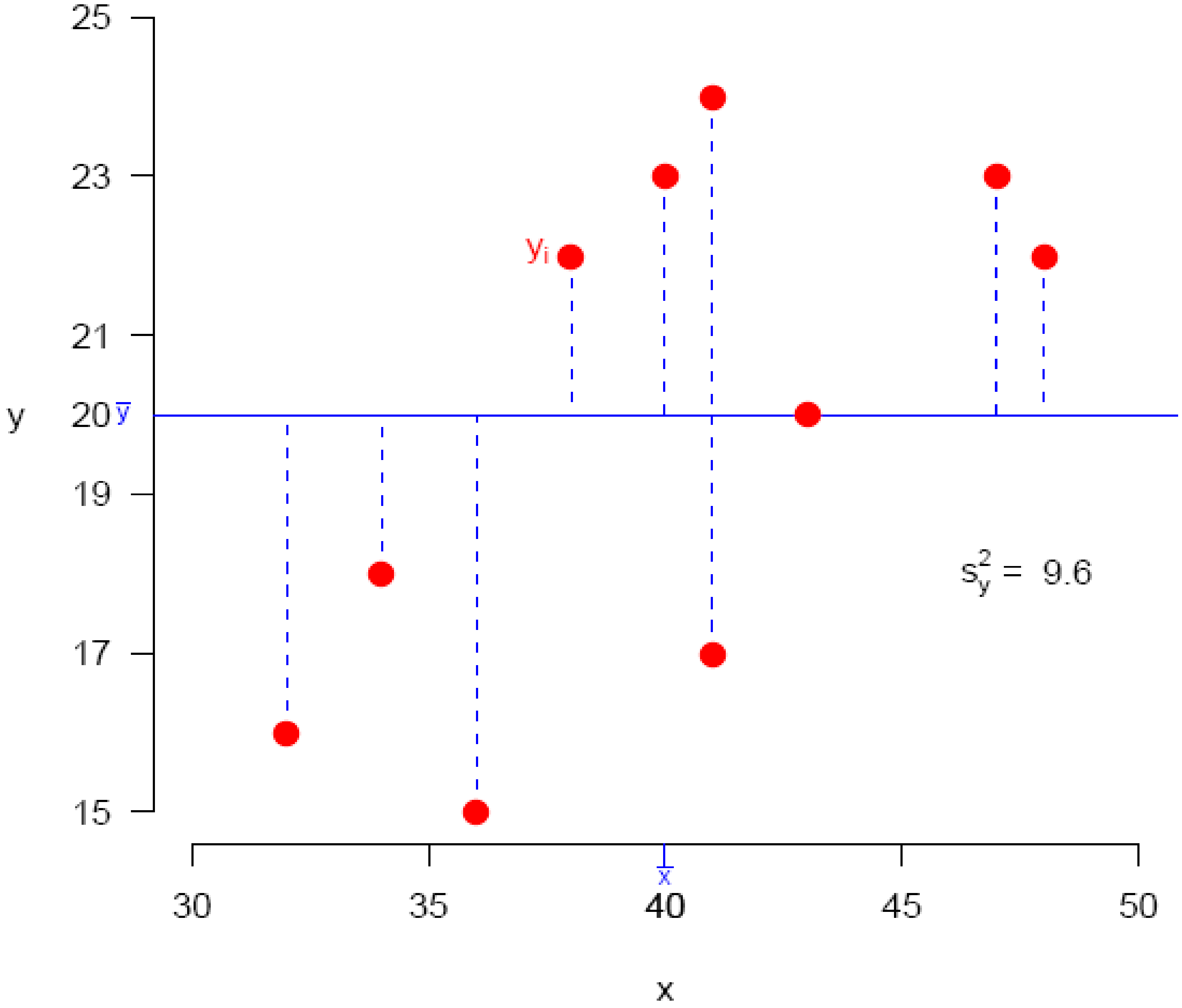


*X*rekiko *Y*ren erregresio lineal bakuna

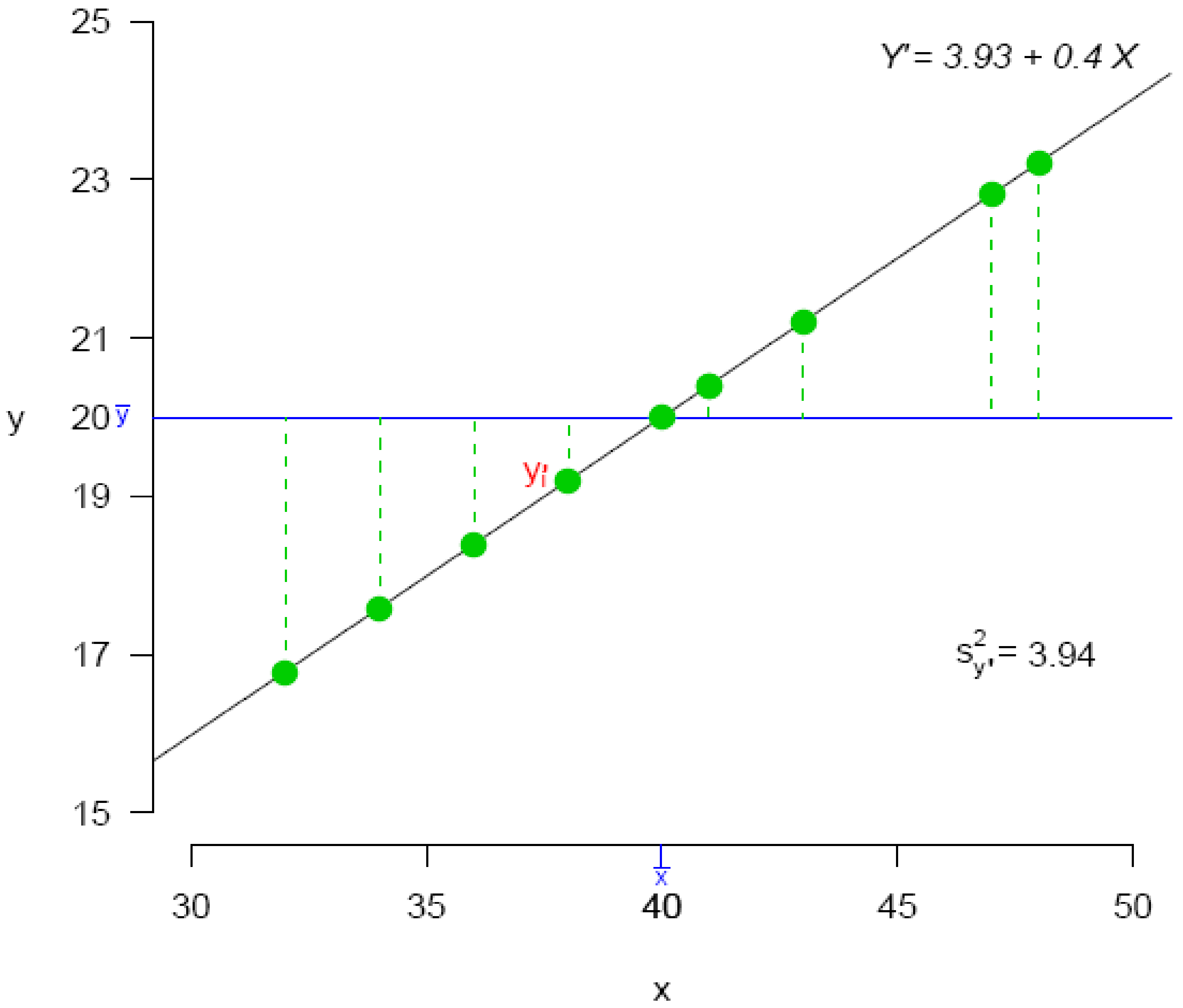
(ebaluazioa)



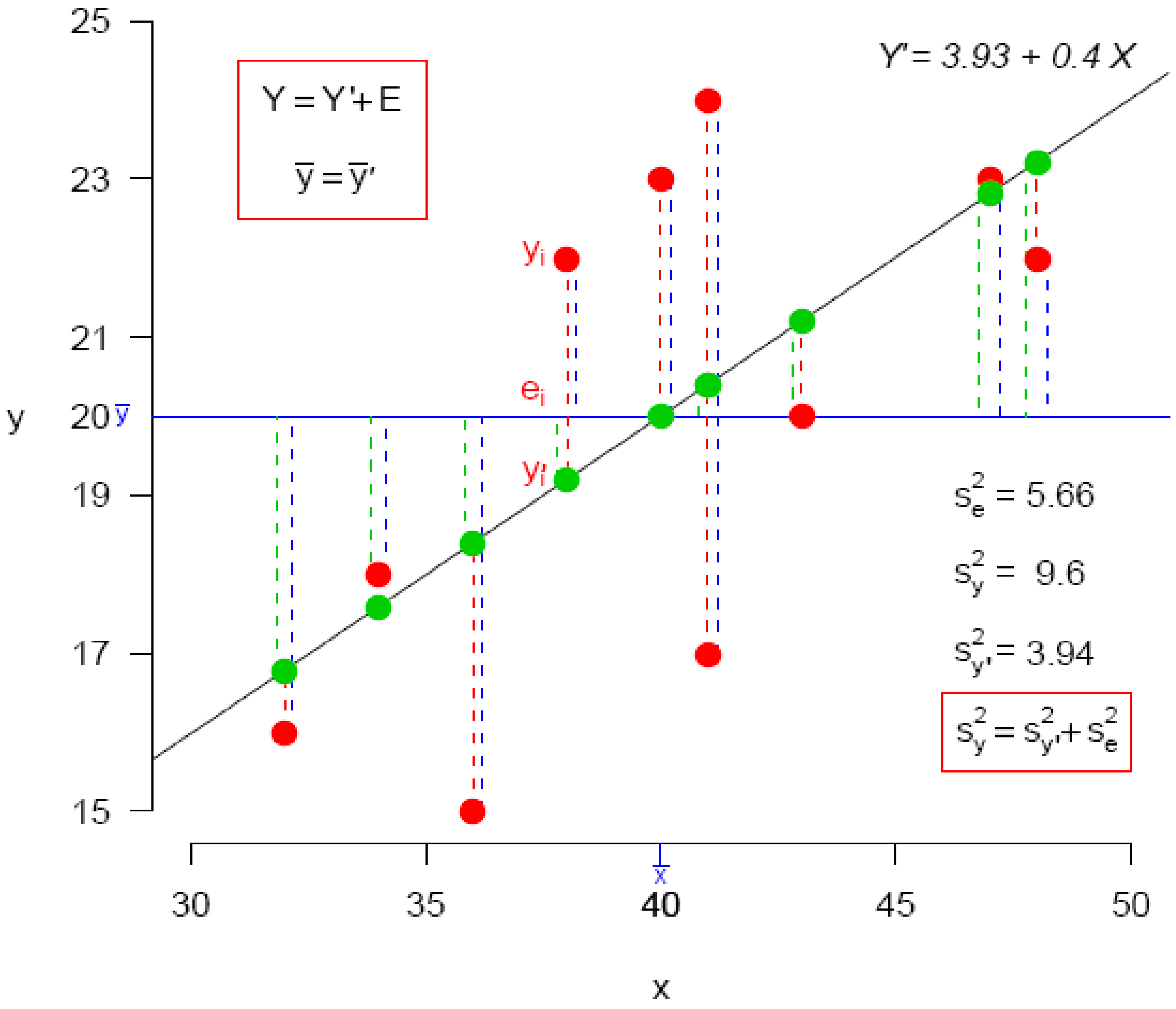
*X*rekiko *Y*ren erregresio lineal bakuna



*X*rekiko *Y*ren erregresio lineal bakuna



*X*rekiko *Y*ren erregresio lineal bakuna



*X*rekiko *Y*ren erregresio lineal bakuna

$$s_y^2 = s_{y'}^2 + s_e^2$$

*X*rekiko *Y*ren erregresio lineal bakuna

$$s_y^2 = s_{y'}^2 + s_e^2$$

Yren BARIANTZA = ERREGRESIOari zor zaion BARIANTZA + ERROREari zor zaion BARIANTZA

Guztirako karratuen batura $\quad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Erregresioaren karratuen batura $\quad SSR = \sum_{i=1}^{n} (y_i' - \bar{y})^2$

Erroreen karratuen batura $\quad SSE = \sum_{i=1}^{n} (y_i - y_i')^2$

$$SST = SSR + SSE$$

*X*rekiko *Y*ren erregresio lineal bakuna

(ebaluazioa)

$s_y^2 = s_{y'}^2 + s_e^2$ $SST = SSR + SSE$

Determinazio-koefizientea: $R^2 = \frac{s_{y'}^2}{s_y^2}$ $R^2 = \frac{SSR}{SST}$

$$R^2 = \frac{s_{xy}^2}{s_x^2 \cdot s_y^2} \qquad R^2 = r^2$$

$$0 \leq R^2 \leq 1$$

$R^2 = 0$ eta $s_{y'}^2 = 0$, orduan $Y'(\omega_i) = \overline{y}$

$R^2 = 1$ eta $s_e^2 = 0$, orduan $Y(\omega_i) = a \cdot X(\omega_i) + b$

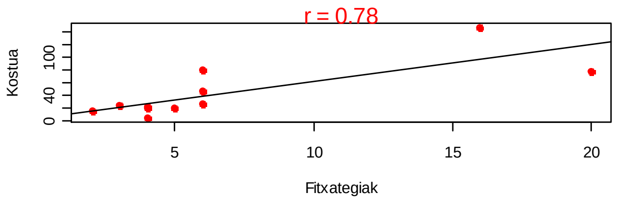
*X*rekiko *Y*ren erregresio lineal bakuna

$$s^2_{y'} = R^2 \cdot s^2_y$$

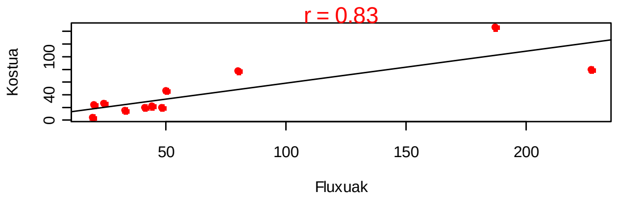
$$s^2_e = s^2_y \cdot (1 - R^2)$$

*X*rekiko *Y*ren erregresio lineal bakuna

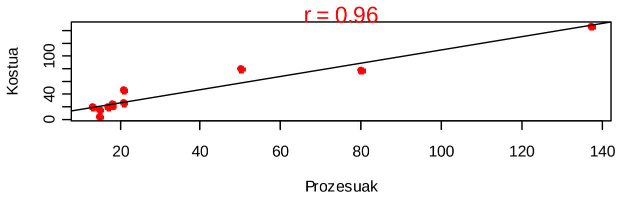
Kalkulatu ***Kostua*** eta beste hiru aldagaien arteko erregresio-ereduen *determinazio-koefizientea*ren balioak



r=0.78



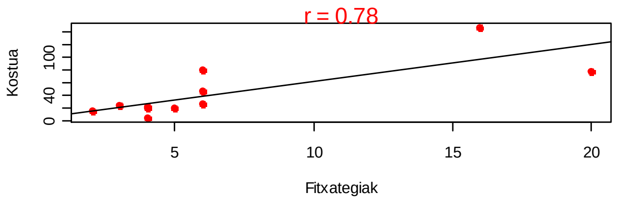
r=0.83



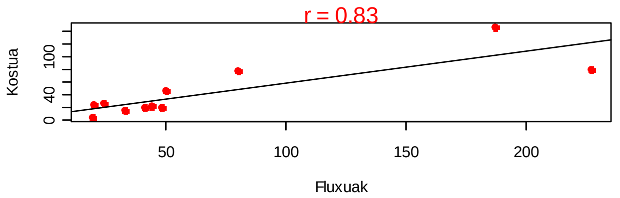
r=0.96

*X*rekiko *Y*ren erregresio lineal bakuna

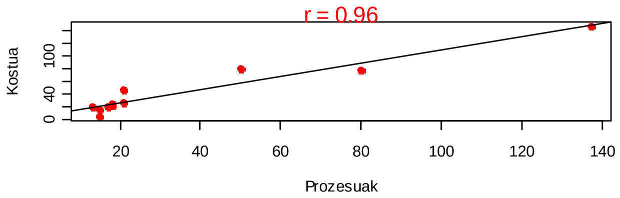
Kalkulatu ***Kostua*** eta beste hiru aldagaien arteko erregresio-ereduen *determinazio-koefizientea*ren balioak



$r=0.78$



$r=0.83$



$r=0.96$

```
> eval(expression(paste("r2 =", round(cor(Kostua,Fitxategiak)^2,digits=2))))
[1] "r2 = 0.61"
> eval(expression(paste("r2 =", round(cor(Kostua,Fluxuak)^2,digits=2))))
[1] "r2 = 0.69"
> eval(expression(paste("r2 =", round(cor(Kostua,Prozesuak)^2,digits=2))))
[1] "r2 = 0.92"
```

***X*ekiko *Y*ren erregresio lineal anizkoitza**

Hobetuko al da ***Kostua***ren erregresioa hiru aldagaiak (***Fitxategiak***, ***Fluxuak***, ***Prozesuak***) linealki konbinatzen badira? Zenbateraino? Hiru aldagaiak beharrezkoak al dira?

*X*ekiko *Y*ren erregresio lineal anizkoitza

Hobetuko al da ***Kostua***ren erregresioa hiru aldagaiak (***Fitxategiak***, ***Fluxuak***, ***Prozesuak***) linealki konbinatzen badira? Zenbateraino? Hiru aldagaiak beharrezkoak al dira?

```
> cor(datuak)
>                 Kostua Fitxategiak   Fluxuak Prozesuak
> Kostua      1.0000000   0.7784743 0.8303919 0.9598421
> Fitxategiak 0.7784743   1.0000000 0.4589818 0.8545609
> Fluxuak     0.8303919   0.4589818 1.0000000 0.7204369
> Prozesuak   0.9598421   0.8545609 0.7204369 1.0000000
```