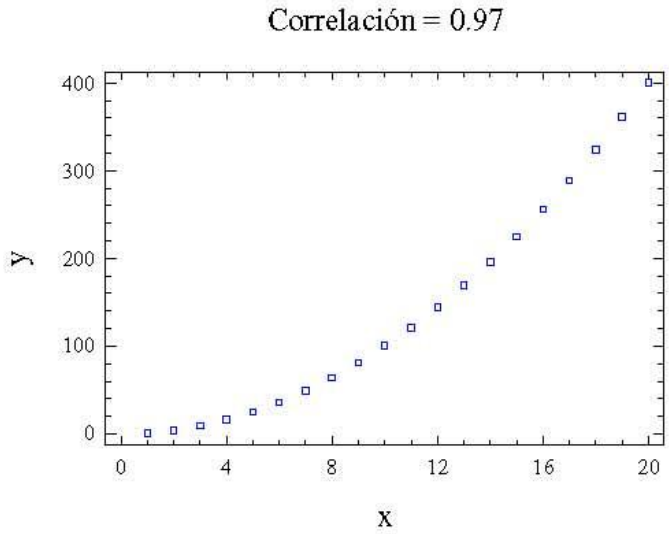
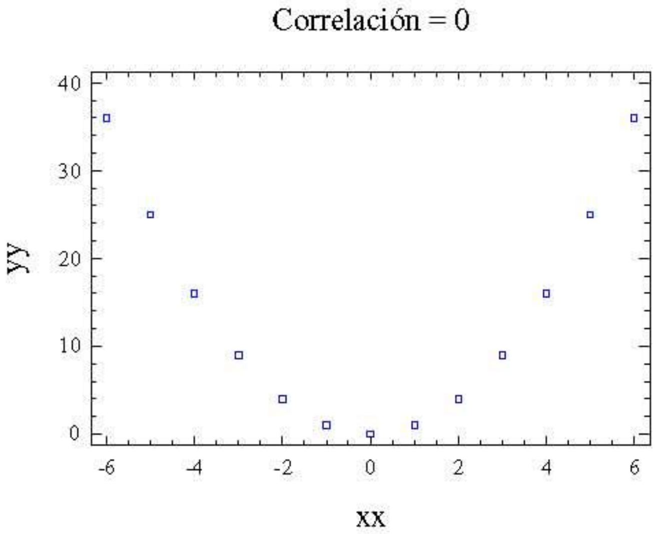
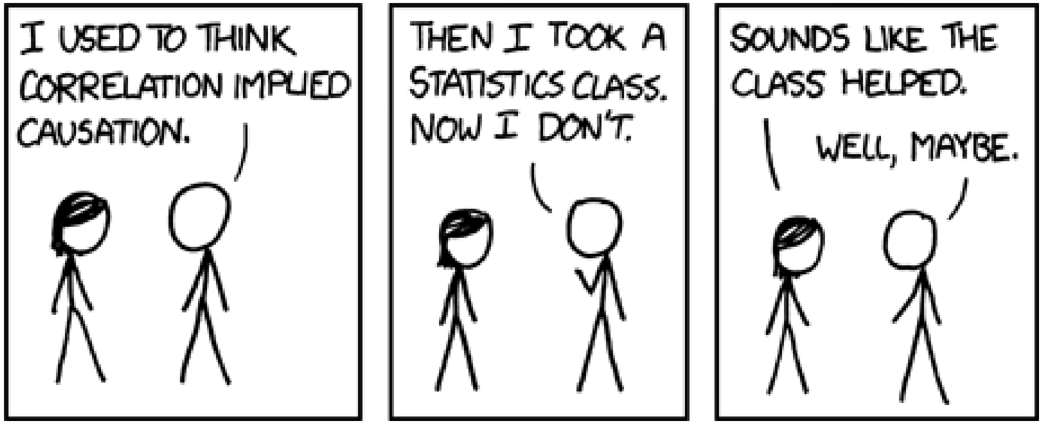
Correlation and non-linear relationships





In both graphs we have set $y=x^2$. A strong, non-linear relationship!

Correlation and causation I


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

A NEW STUDY FOUND COLORADO HAS THE LOWEST OBESITY LEVEL OF ANY STATE.
8/4
YOU KNOW WHAT THAT MEANS?
NO– WHAT?
© 2008 by NEA, Inc.
www.comics.com
FORGET DIETING– WE'RE MOVING TO COLORADO!
© NEA, Inc.

Correlation and causation II

Homer: Not a bear in sight. The Bear Patrol must be working like a charm!

Lisa: That's specious reasoning, dad.

Homer: Why thank you, honey.

Lisa: By your logic, I could claim that this rock keeps tigers away.

Homer: Hmm. How does it work?

Lisa: It doesn't work; it's just a stupid rock!

Homer: Uh-huh.

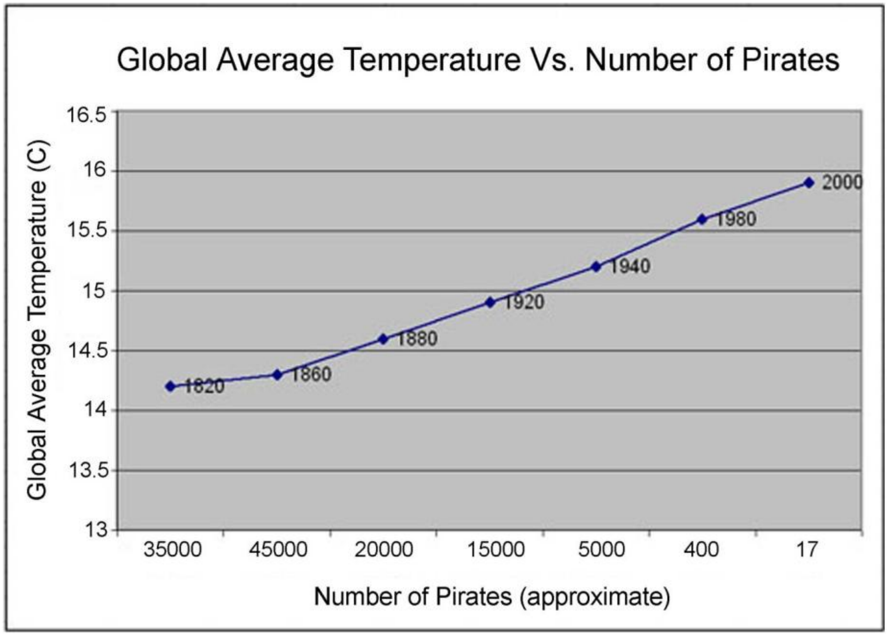
Lisa: But I don't see any tigers around, do you?

Homer: Hmm... Lisa, I want to buy your rock.

Correlation and causation III

What could be the real underlying cause?

STOP GLOBAL WARMING: BECOME A PIRATE



WWW.VENGANZA.ORG

To find more on this in the International Relations context, this video is interesting.

Exercise

A survey of 474 employees was carried out by an multinational company. Among the data gathered were data on salary and years of education. Supposing that Y = Salary and X = years of education

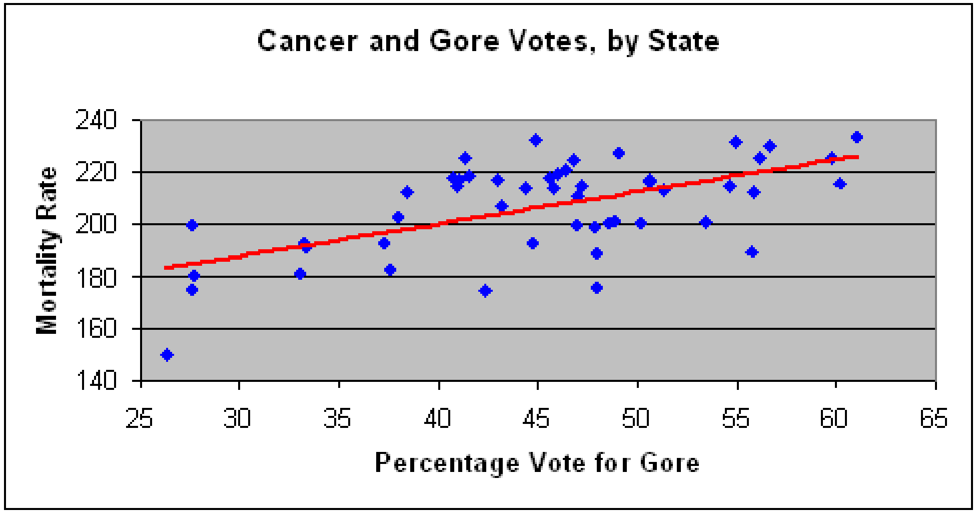
Variance X = 8,305 Variance Y = 290,963 Covariance = 32,471

Mark the correct value of the correlation:

a) -0,53
b) 0,066
c) -0,662
d) 0,662

Exercise

What do you think?

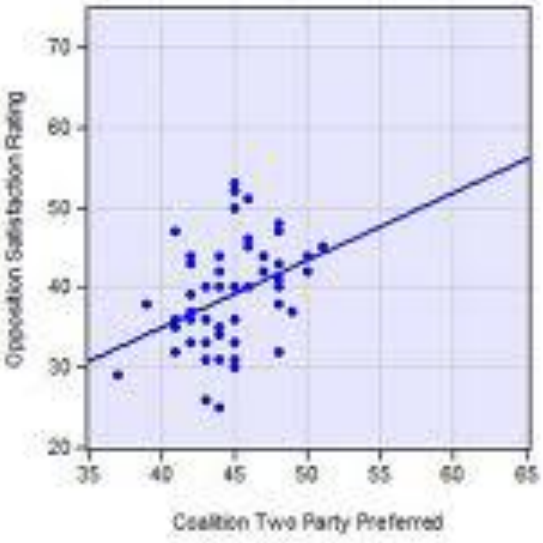


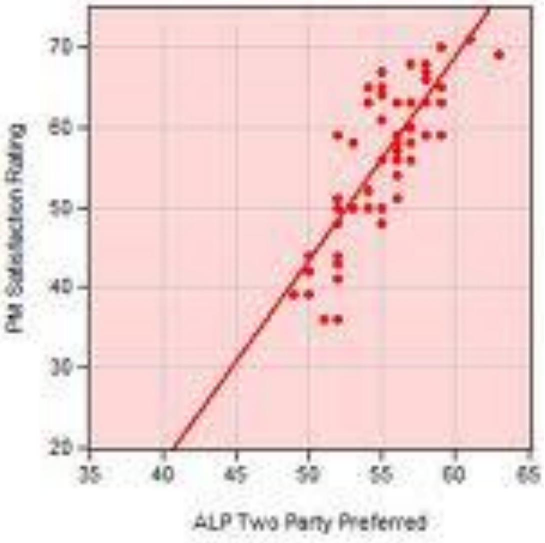
The Hoven study concluded that "[v]oting Democrat is associated with cancer mortality." This is similar to the conclusion of the study **"**Health Insurance and Mortality in US Adults," cited by Democrats in support of their version of health care reform. That latter study concluded that "[u]ninsurance is associated with mortality.“

http://www.americanthinker.com/articles/2010/01/voting_democrat_causes_cancer.html#ixzz3S5qL6pdo

Exercise

The following diagrams show the levels of satisfaction with the party leader and the two party preferred vote in Australia. The diagram on the left hand side is for the opposition party and the diagram on the right hand side is for the government.





Which of the following statements is correct?

a) In both cases, the correlation is negative.

b) The correlation with the two party preferred vote is higher for the opposition party.

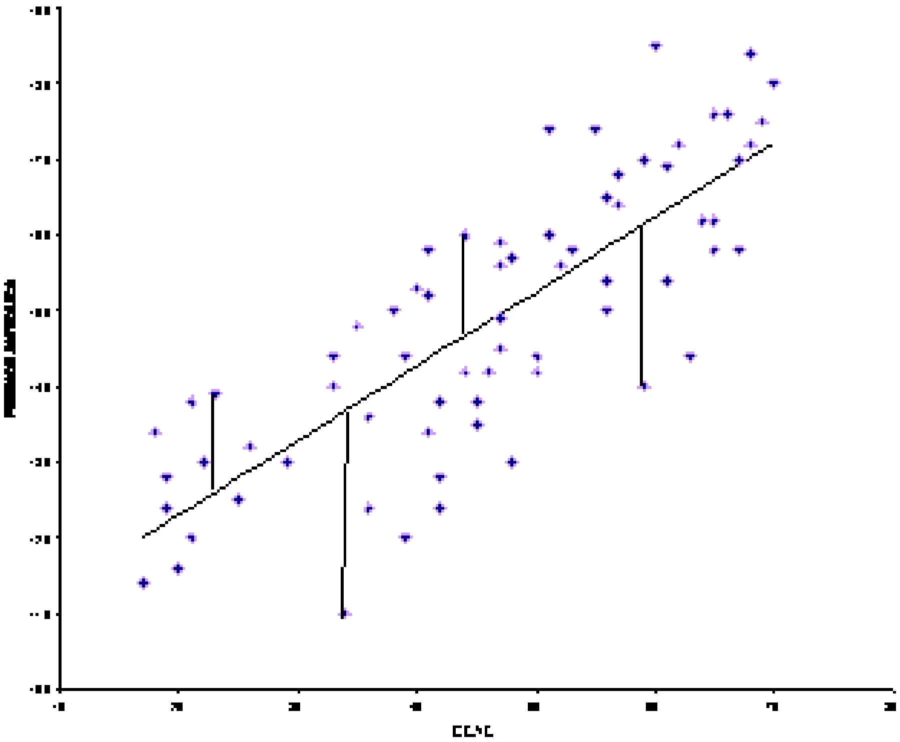
c) The correlation with the two party preferred vote is higher for the government.

d) None of the above.

The regression line

$(x_1, y_1), (x_2, y_2),\ldots,(x_N, y_N)$: N pairs of observed points

We have to find a line: $y = \alpha + \beta x$ which fits our data in “the best possible way”



How do we fit the line?

- We want to predict y given x.
- If we use a line $y = \alpha + \beta x$, then the residuals or prediction errors are $r_i = y_i - \alpha - \beta x_i$ for i = 1,…,N.
- Let's try to minimise the total error.
- Use the least squares criterion: choose the line that minimizes Σr_i^2
- This line is $y = a + bx$

  where b is the slope of the line and a is the intercept:

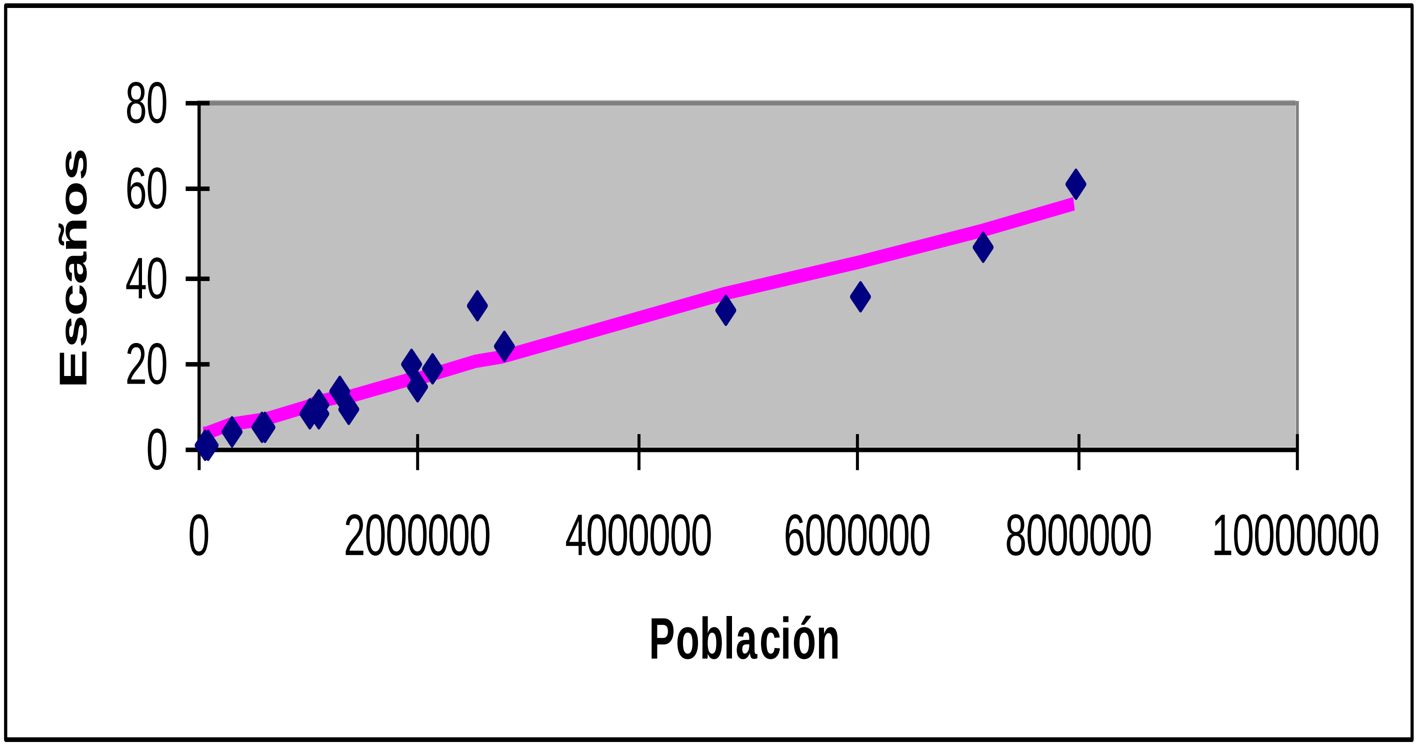
$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

Proof (aagh)

$$
\begin{aligned}
\sum_{i=1}^{N} r_i^2 &= \sum_{i=1}^{N}(y_i-\alpha-\beta x_i)^2\\
&= \sum_{i=1}^{N}(y_i-\bar{y}+\bar{y}-\alpha-\beta x_i+\beta\bar{x}-\beta\bar{x})^2\\
&= \sum_{i=1}^{N}\left(y_i-\bar{y}-[\alpha-\bar{y}+\beta\bar{x}]-\beta[x_i-\bar{x}]\right)^2\\
&= \sum_{i=1}^{N}(y_i-\bar{y})^2+(\alpha-\bar{y}+\beta\bar{x})^2+(\beta[x_i-\bar{x}])^2\\
&\quad -2(y_i-\bar{y})(\alpha-\bar{y}+\beta\bar{x})-2(y_i-\bar{y})\beta(x_i-\bar{x})+2(\alpha-\bar{y}+\beta\bar{x})\beta(x_i-\bar{x})\\
&= Ns_y^2+N(\alpha-a)^2+N\beta^2s_x^2-2(\alpha-a)\sum_{i=1}^{N}(y_i-\bar{y})-2N\beta s_{xy}+2(\alpha-a)\beta\sum_{i=1}^{N}(x_i-\bar{x})\\
&= Ns_y^2+N(\alpha-a)^2+N\beta^2s_x^2-2N\beta s_{xy}\\
&= Ns_y^2+N(\alpha-a)^2+N\beta^2s_x^2-2N\beta s_{xy}+N\left(\frac{s_{xy}}{s_x}\right)^2-N\left(\frac{s_{xy}}{s_x}\right)^2\\
&= Ns_y^2+N(\alpha-a)^2+Ns_x^2\left(\beta-\frac{s_{xy}}{s_x^2}\right)^2-N\left(\frac{s_{xy}}{s_x}\right)^2\\
&= N(\alpha-a)^2+Ns_x^2\left(\beta-b\right)^2+Ns_y^2-N\left(\frac{s_{xy}}{s_x}\right)^2
\end{aligned}
$$

y se minimiza esta función con $\alpha = a$ y $\beta = b$.

Seats and population: The fitted regression line



Excel Output

	Coeficientes
Intercepción	2,692069443
Variable X 1	6,68437E-06

The fitted line is y = 2,69+0,0000069x

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,96372808
Coeficiente de determinación R^2	0,928771813
R^2 ajustado	0,92458192
Error típico	4,544275594
Observaciones	19

How do we predict the seats is a community of 1000000 people?

And in a community with no people?

Does this prediction make sense?

Residual analysis I: residual mean and variance

The mean of the residuals is 0.

$$
\begin{aligned}
\sum_{i=1}^{N} r_i &= \sum_{i=1}^{N} (y_i - a - bx_i) \\
&= \sum_{i=1}^{N} (y_i - a - bx_i) \\
&= \sum_{i=1}^{N} (y_i - \bar{y} + b\bar{x} - bx_i) \\
&= \sum_{i=1}^{N} (y_i - \bar{y}) - b \sum_{i=1}^{N} (x_i - \bar{x}) \\
&= 0
\end{aligned}
$$

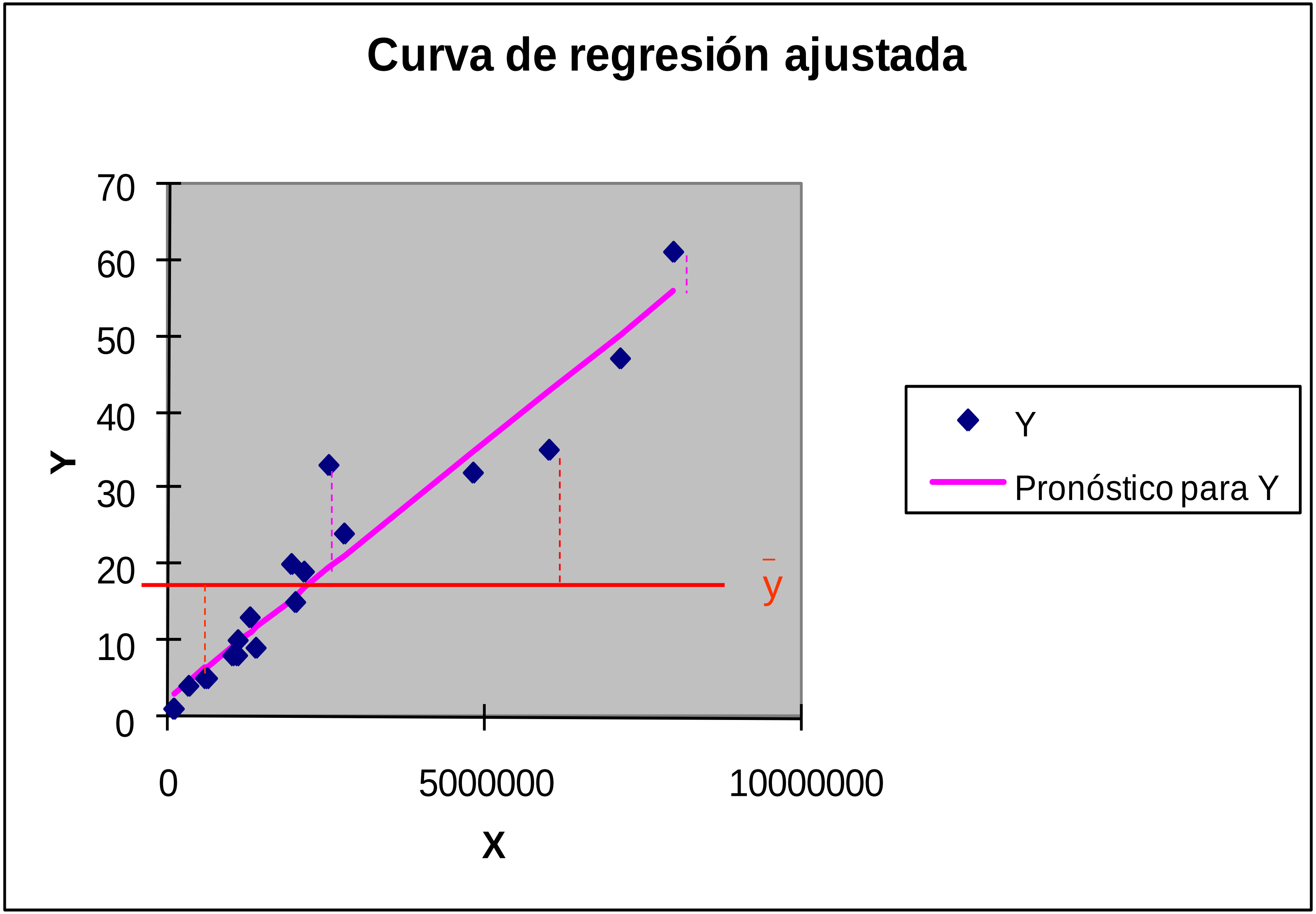
And the variance can be calculated as

$$\begin{aligned}\sum_{i=1}^{N} r_i^2 &= N s_y^2 - N\left(\frac{s_{xy}}{s_x}\right)^2 \\ &= N s_y^2 \left(1 - \frac{s_{xy}}{s_x s_y}\right)^2 \\ &= N s_y^2 \left(1 - r_{xy}^2\right)\end{aligned}$$

Luego, la varianza residual es

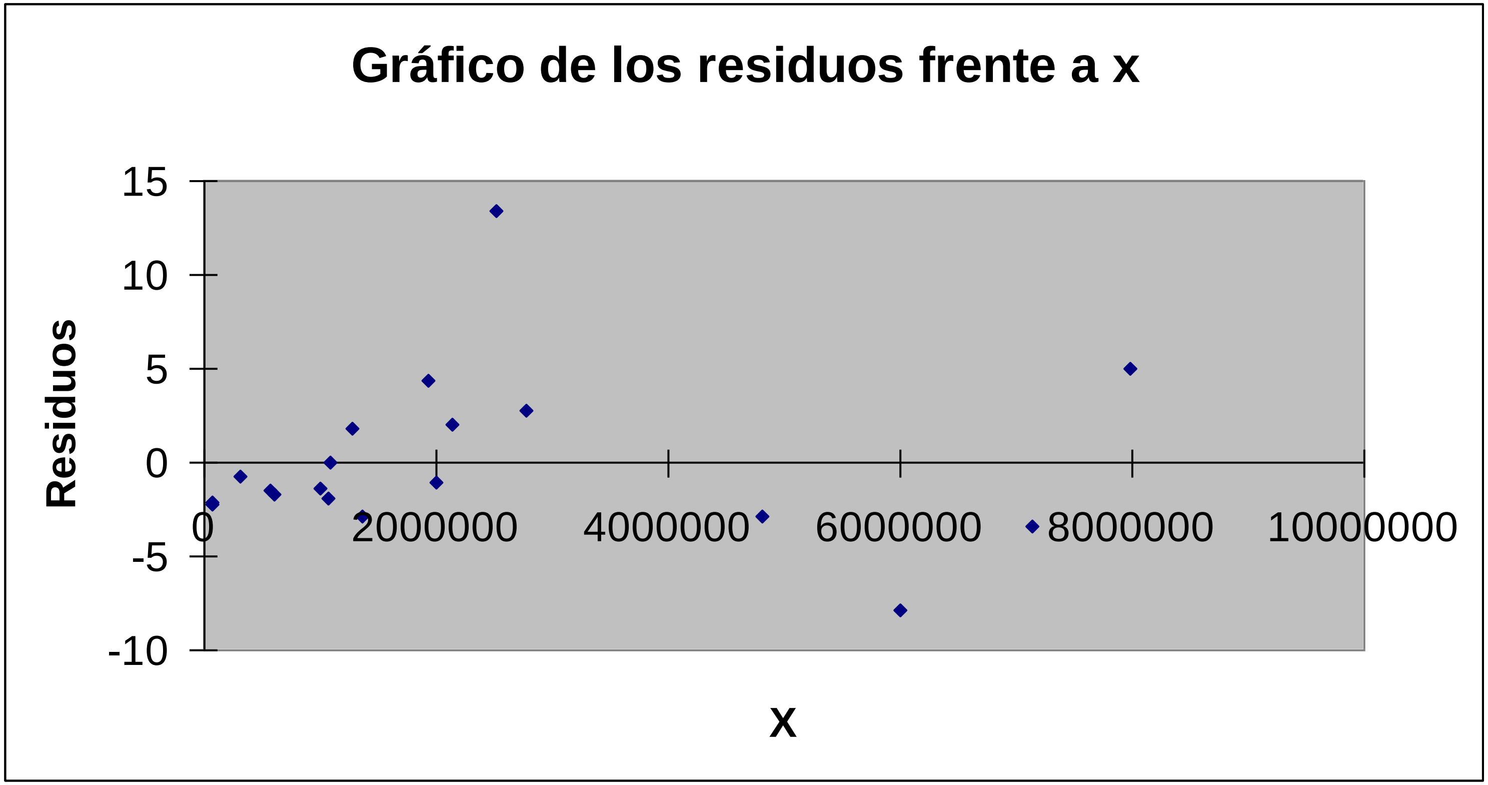
$$s_r^2 = \frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = s_y^2(1 - r_{xy}^2).$$

How do we interpret this?


Curva de regresión ajustada
70
60
50
40
30
20
10
0
Y
0
5000000
10000000
X
$\bar{y}$
Y
Pronóstico para Y

Residual analysis II: graphs

If the regression line fits the data well, the residuals should look like “random noise” with no relation to x or y.



Does this fit look good?

Example

The table shows the Gross National Product per head in US dollars in 2008 and 2009 for the G8 countries.

	GNP 2008	GNP 2009
Country	x	y
Canada	42030	39217
France	45981	42091
Germany	44471	39442
Italy	38309	34955
Japan	38443	39573
Russia	11339	8874
UK	43088	35728
USA	46716	46443

The covariance between the two variables is 116000000 and the correlation is 0,974. The Libyans prefer to measure GNP in Libyan dinars. The dollar dinar Exchange rate is (approximately) 1 dollar = 2 dinars.

Measuring the GNP per head in Libyan dinars, which of the following options is correct?

a) Both the covariance and the correlation do not change.
b) The correlation is 0.2475 and the covariance does not change.
c) The covariance is 464000000 and the correlation doesn’t change.
d) Both the covariance and the correlation change to a quarter of their previous values.

Example

The following table shows information about the daily sales of newspapers for each 1000 inhabitants of 8 Spanish Communities and the economic production of the community based on the PIB (Producto Interiór Bruto) per resident .

PIB	8.3	9.7	10.7	11.7	12.4	15.4	16.3	17.2
Sales	57’4	106’8	104’4	131’9	144’6	146’4	177’4	186’9

Suppose a linear relation between these variables, we obtain the following regression line which explains the number of papers sold per 1000 inhabitants in terms of the PIB per resident in 1000’s of euros:

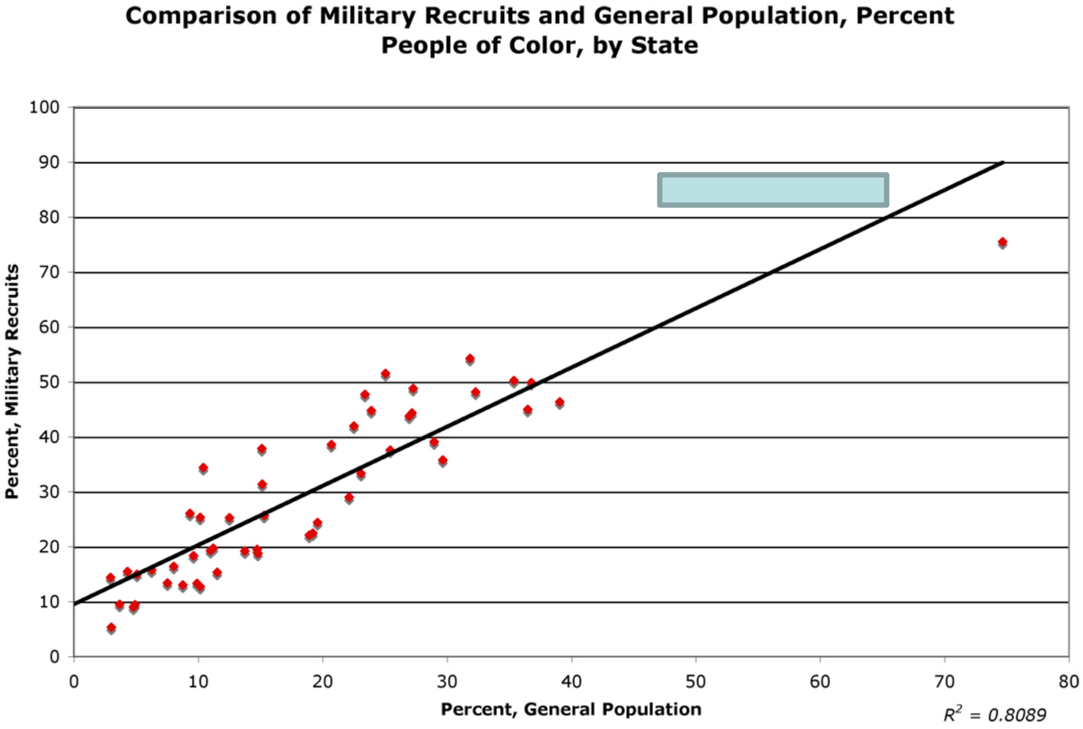
$$y= -23.55 + 12.23x$$

What would be the predicted sales in a community with PIB per resident equal to 15.000 euros?

a) 159.9 examples
b) 159.9 examples for each 1000 inhabitants
c) 183.430 examples
d) 183.430 examples for each 1000 residents

Example

A US newspaper is carrying out a study on racism in the US army. They have calculated the following scatterplot shows the percentages of coloured military recruits (y) against the general population size (x) for various US states.



Which one of the following regression lines is correct?

a) y = 1.08x
b) y=9.55-1,08x
c) y=9.55+1,08x
d) y=-9.55-1,08x

Example

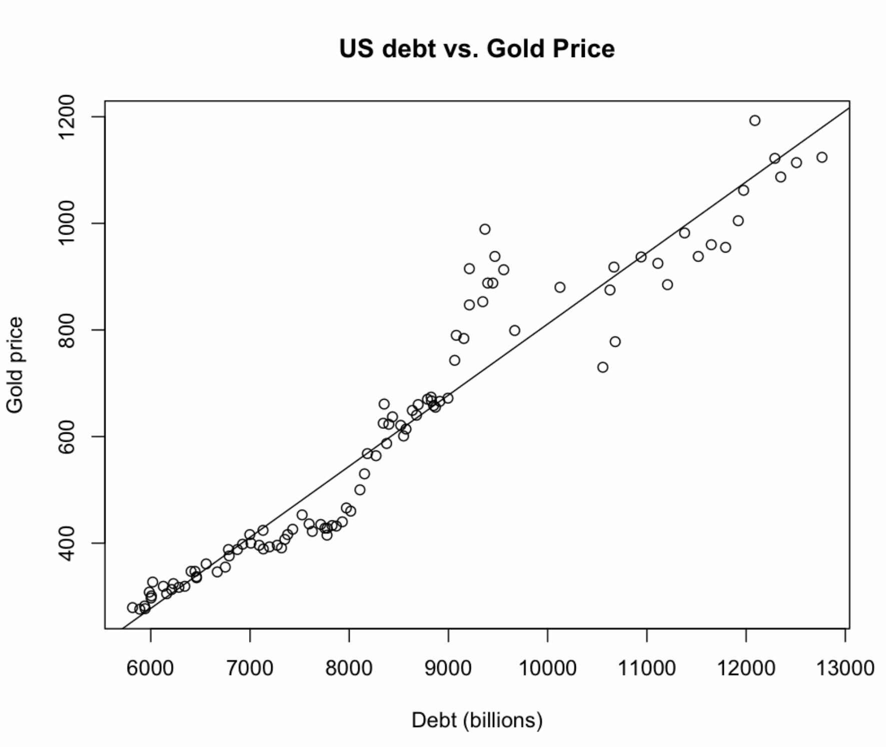
The diagram shows the level of US debt as a function of the gold price.

The linear regression formula (without the error term) is:

GOLD PRICE (nominal) = -522.86 + (0.1334 * US-debt-in-billions)

If the US debt is $19000 billions, what would you predict the gold price to be?

a) 2011.74
b) 3057,46
c) 2933,14
d) -520.3254



Do you think it is reasonable to use the regression line to make your prediction in this case? If not, why not?